THE UNIVERSITY OF HULL

AN EVALUATION OF SELECTED ESTIMATION METHODS FOR THE PROCESSING OF DIFFERENTIAL ABSORPTION LIDAR DATA

being a Thesis submitted for the Degree of

Doctor of Philosophy

in the University of Hull

by

Andrew Layfield, B.Sc., M.Sc.

March 1987

Summary of Thesis submitted for Doctor of Philosophy

by Andrew Layfield

on

AN EVALUATION OF SELECTED ESTIMATION METHODS FOR THE PROCESSING OF DIFFERENTIAL ABSORPTION LIDAR DATA

This work examines the application of selected estimation methods to path integrated direct detection CO_2 lidar data, with the objective of improving the precision in the estimates of the log power, and log power ratios. Particular emphasis is given to the optimal estimation techniques of Kalman filtering theory, and to the consequent requirements for system and measurement model identification. A dual wavelength system was designed and constructed, employing two hybridised TEA lasers, a co-axial transceiver, and direct detection.

Over a period of several months, a database of differential absorption measurements was accumulated, each consisting of 10,000 dual wavelength lidar returns. Various wavelength pairs were used, including those recommended for the monitoring of H_2O, CO_2, NH_3 and C_2H_4. A subset of this database is used to evaluate the above mentioned estimation methods. The results are compared with simulated data files in which it was possible to control precisely process models which are believed to form an approximation to the real processes latent in the actual lidar data.

To Tina

ACKNOWLEDGEMENTS

I would like to thank Dr. B. J. Rye for his continued guidance and encouragement throughout the work presented here, Dr. E. L. Thomas for his earlier support, and the Royal Signals and Radar Establishment who contributed the CASE award (particularly the late Dr. Ken Hulme). Many thanks to Pete Fountain and Pete Seviour whose technical assistance proved to be invaluable, and to Dr. S. E. Taylor for his reading of some of the chapters and valuable comments.

Particular gratitude is due to Helen Lythgoe for completing all of the typing to such excellent standards within the "eleventh hour" I had reserved for its completion.

Finally, in dedicating this thesis to Kristina Gorman I am acknowledging her unfailing support and help, without which it could never have been produced.

CONTENTS

CHAPTER 1.0

INTRODUCTION

1.1 LIDAR Systems

The application of the optical counterpart to Radar, now generally referred to by its acronym LIDAR for "Light Detection and Ranging", has encompassed many of the developments in tunable laser sources. Transmitted wavelengths range from the ultra-violet through visible to the infra-red, and have been used for a wide variety of measurements including temperature, humidity, wind velocity, cloud base heights, dust and aerosol distributions, variations in the major constituents of the Earth's atmosphere and, of course, pollutant or trace gas concentrations [1-16].

In measuring trace gases, some of these systems use topographic targets or retro-reflectors and are therefore restricted to path integrated measurements. Other systems, however, use the back-scattered radiation from one or more of the atmospheric constituents to provide the return signal. Both Raman and resonance scattering have been employed to indicate directly the presence and concentration of pollutant molecules. Alternatively, similar information can be obtained indirectly using the radiation scattered from naturally occurring aerosols. Inherent in the design of all atmospheric scattering systems is the ability to range resolve the parameters outlined above.

"Single ended" or "monostatic" systems confine all the necessary equipment to a single location, yet enable measurements to be made over an extended volume of air space, often with dimensions of cubic kilometres. As such they offer an alternative to in-situ sensors which are

capable of measuring gas concentrations only at a single location. This is important in situations where measurements need to be made in places physically inaccessible or else hazardous to both equipment and operator. Advantages are also found in applications where the equipment must be portable; examples here include the various ground mobile and air-borne systems, and the LIDAR payloads originally intended for launch by the NASA Shuttle in the mid-1980's [17,18].

1.1.1 The Hull LIDAR System

The LIDAR system developed at Hull University used the differential absorption of two CO_2 laser pulses, transmitted almost simultaneously along the same optical path in the atmosphere, as the basis of a ratioing technique to determine the concentration of an absorbing molecule. The output from one laser is absorbed by the molecular species of interest whilst the other, tuned to a slightly different frequency, is subject to less absorption (ideally non at all). Acronyms classifying this type of system will therefore often include the component "Differential Absorption". For example, if atmospheric backscattering is employed to provide the return signal, then "DIAL" (Differential Absorption Lidar) is used. The term "Lidar", however, has come to be used as a generic description for any long propagation path system but, by definition, should be reserved for systems capable of range resolved measurements.

A full description of the system configuration will follow in Chapter 2 but, for the purposes of the analysis conducted in this work, the radiation collected by the receiving telescope was detected directly rather than by the more sensitive, but significantly more complex method of coherent detection.

1.2 Pollution and Trace Gas Absorption Lines of Interest

Table 1.1 gives a representative list of pollutants and trace gases currently of interest for meteorological reasons, for investigations in atmospheric science and transport studies, or for their relevance as possible detriments to the environment. Strictly speaking, water vapour and CO_2 are not pollutants or trace gases but important atmospheric constituents with concentrations which can vary both diurnally and with changes in geographical location.

Ozone is also a naturally occurring atmospheric component but most of it is concentrated within a layer two or three miles thick, at some 15 - 20 miles above the surface of the Earth. It is, however, toxic in concentrations as low as 100 ppb causing respiratory and visual problems [19]. Atmospheric transport processes can cause this stratospheric ozone to descend to lower altitudes within the troposphere. Other natural sources at lower altitudes include lightning and volcanic eruptions. Man-made additions tend to originate in industrial waste gases, particularly hydrocarbons and nitrous oxides (NO_x) which react in the presence of sunlight to form ozone. Typical sea-level concentrations range between 10 and 30 ppb but levels as high as 500 ppb have also been observed [19].

Gases such as Sulphur Dioxide and Ammonia occur naturally in low background concentrations ($\lesssim$ 0.01 ppm) but near industrialised areas the concentrations can be much higher. Large numbers of deaths have been attributed indirectly to SO_2 in polluted areas in which the concentrations reached more than 1.0 ppm [19]. Accelerated ageing is also reputed to result from exposure to this gas [20]. Ammonia too is directly harmful to man but is under investigation because of its important role in the production and destruction of the oxides of Nitrogen in the stratosphere and troposphere [19]. Such processes can be potentially damaging if they

reduce the concentration of ozone in the stratosphere since this layer acts as a shield, absorbing most of the harmful ultra-violet radiation from the sun.

Numerous hydro-carbons exist as forms of pollutants in the atmosphere and are capable of interacting in a diverse number of ways with other atmospheric molecules. Candidates for investigation include ethylene (C_2H_4), again an industrial by-product, and also propane (C_3H_8) which can occur in background levels ranging from a few ppm to extreme concentrations of 40 ppm [1]. Similar interactions involving the various types of Freon are also the subject of investigation [21].

Table 1.2 (extracted from [1]) gives the composition of "clean" dry air near sea level, including the background concentrations of a number of minor constituents described briefly above.

1.3 The Application of CO_2 Lasers

1.3.1 Atmospheric Transmission

An important criterion in selecting the transmitted wavelengths for a laser radar system is the atmospheric attenuation over the various wavelength ranges of interest (i.e. available within the current developments of laser technology). Energy dissipation from an optical beam propagating in the atmosphere can be attributed primarily to the following loss mechanisms:

(1) Molecular absorption

(2) Molecular scattering

(3) Particulate scattering

(4) Particulate absorption

(5) Atmospheric turbulence.

(1) Molecular absorption (denoted, α_M) is generally the result of the super-position of the absorption spectra of more than one molecular species. Fig. 1.1 is a low-resolution representation of the absorption spectra of all the atmospheric gases, over a wavelength range from ~0.1 - 100 μm, at two altitudes above the Earth's surface. The upper curve illustrates the attenuation of solar radiation reaching the surface whilst the lower curve is for the radiation reaching an altitude of 11 km. Contributions due to the various gases listed in Table1.2 are also indicated. They clearly indicate the presence or absence of "windows" over various wavelength regions from the ultra-violet to the infra-red with nearly complete absorption occurring beyond the limits of 0.2 and 20 μm. At short wavelengths this absorption is due to oxygen and ozone, and at longer wavelengths to water vapour.

These curves are somewhat misleading, however, since higher resolution spectra would in fact reveal each absorption band and each "window" as consisting of thousands of separate absorption lines. Each region is therefore neither completely opaque or completely transparent. This is fortunate because some of the narrower windows happen to coincide, or partly overlap, with the emission spectra of various lasers. Determining the transparency of these windows at high resolution often requires empirically derived results.

Of particular interest to the application of CO_2 lasers is the window existing in the 9 - 11 μm region, not resolved in Fig. 1.1, but partially represented in Fig. 1.2 complete with details of the fine spectra [22].

Transmittance values are calculated over a range of 10 km for a horizontal sea level path and, for the 10 - 9 μm region

illustrated, vary from negligible absorption to approximately 70%. The spectrum actually depicts a convolution of strong atmospheric absorption features with an almost constant background absorption known as the "water vapour continuum". This is thought to be due to water molecules forming themselves into more complicated basic units known as "dimers" and "trimers" [23].

Also evident in Fig. 1.2, appear to be the P and R branches of the absorption spectra of CO_2. A comparison with Fig. 1.3, giving the emission spectra of CO_2 alone, will reveal the spectral location of these features. As indicated in Table 1.2, CO_2 occurs as a natural constituent of the Earth's atmosphere and some absorption can be expected. However, since the stimulated emission lines of a CO_2 laser do not involve transitions directly from the ground state, absorption by atmospheric CO_2 will fortunately be limited. Strong absorption features due to other gases also exist, however, and specific CO_2 lines will not always coincide with spectral intervals of relatively high transmittance.

(2) Molecular scattering is usually treated using the Rayleigh scattering theory for gas molecules which relates the extinction coefficient α to the number of scatters per unit volume (n_s), the refractive index of the gas (n) and the wavelength of the incident radiation via expressions of the form [1]

$$\alpha_R(\lambda) = [8\pi^3(n^2-1)/3n_s\lambda^4][(6 + 3\delta)(6 - 7\delta)] \qquad (1.1)$$

where δ is a depolarisation factor. Important here is the λ^{-4} dependence of α_R on wavelength which is obviously advantageous if the laser wavelengths used are in the infra-red rather than visible

or ultra-violet. Fig. 1.4 compares values of α_R over the wavelength range 0.2 to 20 μm [24]. Differences of many orders of magnitude become apparent between the 9 -11 μm range of CO_2 lasers and wavelength below 1 μm.

(3) Treatments of particle scattering often include the general attenuation due to both scattering and absorption of laser radiation by a diversity of air-borne particles. Although the early theoretical treatments were pioneered by Mie in 1908 [1] many subsequent treatments exist [25] because of its significance in all investigations of atmospheric optics. "Mie scattering", as it is now known, is highly dependent on particle size, wavelength and particle distribution with no straightforward relationship existing between the Mie scattering coefficient α_{MIE} and wavelength (unlike Rayleigh scattering). The results of numerical computations are available [20] for various classifications of aerosol distribution but, in general, losses due to particle scattering have a tendency to decrease with increasing wavelength.

(4) Particulate absorption manifests itself in a component term in the expression for particle scattering. Parameters are represented by complex variables, including a complex refractive index, and it is the imaginary terms which yield the contributions due to absorption.

These various extinction parameters can be combined into a single atmospheric attentuation coefficient,

$$\alpha_A = \alpha_M + \alpha_R + \alpha_{MIE} \text{ (scattering, absorption)} \qquad (1.2)$$

and are related to the intensity of the optical radiation propagating in the atmosphere via the differential expression

$$dI(\nu) = - I(\nu)\, \alpha_A(\nu,z)dz \tag{1.3}$$

which is known as either Bouguer's Law (or Beer's Law or Lambert's Law, or the Beer-Lambert Bouguer (BLB) Law). The direction of propagation is z and $dI(\nu)$ is the change in intensity after the radiation has passed through a medium layer of depth dz. In integral form this becomes

$$I(\nu) = I_o(\nu) \exp \left[- \int_o^z \alpha_A(\nu,z)dz \right] \tag{1.4}$$

where I_o is the initial radiation intensity, or

$$I(\nu) = I_o(\nu) \exp \left[- \alpha_A(\nu)z \right] \tag{1.5}$$

if the medium is homogeneous. In Chapter 3.0 this relationship will be incorporated into a general expression for LIDAR systems, used to predict radiation intensities after two-way propagation paths between source and remote target.

(5) Atmospheric turbulence effects redistribute the energy of optical beams and cannot be considered simply in terms of attenuating coefficients, but must be statistically interpreted in terms of the mean and variance of the resultant intensity distribution after a given propagation length in the turbulent medium. For example, the variance of the logarithm of intensity, assuming a point source and a point receiver, may be given by empirically derived expressions of the form [e.g. ref. 38],

$$\sigma^2_{\ln I} = f\left(C^2,\ K^{7/6},\ L^{11/6}\right) \tag{1.6}$$

where C_n^2 is the atmospheric structure constant, or scintillation index, characterising the strength of the turbulence, $k = \frac{2\pi}{\lambda}$ and L is the atmospheric path length. The relevance of turbulence will be discussed in more detail later in Chapter 4.0, but of significance here is the inverse relationship between wavelength and variance; again, advantages are incurred by operating at longer wavelengths.

1.3.2 Coincidence of CO_2 Emission Lines with Trace Gas and Pollutant Absorption Spectra

Table 1.1 presents a list of specific gases of interest accessible with CO_2 wavelengths, together with the recommended measurement and reference lines to be used in comparing absorption. These, however, may not represent the optimum choice because interference from other atmospheric gases can modify the differential absorption coefficient, important in determining molecular concentrations and associated sensitivities (discussed in Chapter 5.0). Included in Table 1.1 are certain pollutants which would seem to be best detected using different lasers such as HeNe, Carbon Monoxide or even isotopic CO_2 lasers (only the measurement lines are shown). The majority, however, given the availability of all known discrete laser transitions, have suitable absorption cross-sections coinciding with the line emissions of ordinary CO_2 lasers.

CO_2 lasers, therefore, have obvious advantages but the problem of the discrete tuning nature of gas lasers at atmospheric pressures and below, means that coincidences between emission and absorption spectra are always fortuitous and seldom ideal. In some cases, for example

sulphur dioxide, the match is not as good, resulting in reduced sensitivity. The advent of reliable, continuously tunable lasers may eventually optimise the situation but at the time of writing these are still under development [20,26,27,34,35,36,37].

1.4 Project Objectives

The investigations planned at the outset of this project were directed at the ultimate objective of remote measurement of trace gases with accuracies approaching a few parts per billion (ppb). A primary requirement was the design and construction of a dual wavelength laser system capable of simultaneously generating two single-transverse-mode pulses of at least 100 mJ, and at relatively high pulse repetition rates. An existing telescopic transmitter and receiver configuration was used and integrated into a system which also included (a) the dual laser source, (b) the optics used to simultaneously couple the two output pulses into a single optical transmission path, and (c) the optical components necessary for the dual direct detection of two signals received simultaneously or with a given time delay.

To obtain differential absorption measurements it was necessary to completely re-configure the original single wavelength system, successfully used in demonstrating the application of coherent lidar to wind velocity measurement [28,31,32,33]. A more stable optical system was constructed in order to avoid the frequent and time-consuming alignment procedures found necessary when using the earlier system. The extra stability was also essential for the increase in optical system complexity which accompanied the inclusion of a dual wavelength laser source.

Originally, attempts to solve the problem of how to improve measurement sensitivity tended to focus on "system hardware" innovations which would increase the number of statistically independent return power samples obtained per unit time. Direct detection of hard target returns would permit only increased laser rep rates, augmented (where practical) with aperture averaging, but coherent detection of atmospheric returns made feasible more exotic techniques. Examples include the use of short duration laser pulses and spatial averaging within gated range cells, or "frequency diversity" in which a long cavity laser system could be used to increase the number of longitudinal modes transmitted and hence, in principle, increase the number coherently detected.

In general, it tended to be implicitly assumed that the absorption (or backscatter) characteristics of the atmosphere remained approximately stationary during the course of the measurements so that each temporally sequential sample was independent and simple averaging techniques could be applied. Evidence was available, however, that these characteristics were, in general, not stationary and that a trend often existed which invalidated the simple averaging techniques [30]. More sophisticated estimation techniques are therefore required which are capable, not only of "tracking" the time dependent qualities of interest as accurately as possible, but also of providing an indication of the precision of the estimates.

It is the objective of this work to examine the application of alternative estimation methods to path integrated direct detection lidar data obtained with the Hull University lidar system, and particular attention will be paid to the optimal estimation techniques of Kalman filtering

theory. Over a period of several months towards the end of 1984 a database of measurements was built up, with each measurement consisting of 10,000 dual wavelength lidar returns. Selected files will be used to evaluate these estimation methods and the results will be compared with simulated data files in which it was possible to control precisely process models which are believed to form an approximation to the real processes latent, actual lidar data.

REFERENCES

1. HINCKLEY, E.D. (1976). Laser Monitoring of the Atmosphere. Springer Verlag, Berlin.

2. BOWMAN, M.R. et al. (1970). The Radio and Electronic Engineer 39(1), pp. 29-32.

3. ROTHE, K.W. et al. (1974). Applied Physics 3, pp. 115-119.

4. MURRAY, E.R. et al. (1976). Appl. Phys. Lett. 28(9), pp. 542-543.

5. MURRAY, E.R. and VAN DER LAAN, J.E. (1978). Appl. Opt. 17(3), pp. 814-817.

6. MURRAY, E.R. et al. (1978). Appl. Optics 17(2), pp. 296-299.

7. MURRAY, E.R. (1978). Optical Engineering 17(1), pp. 30-38.

8. MURRAY, E.R. (1977). Optical Engineering 16(3).

9. CRUICKSHANK, J.M. (1979). Appl. Optics 18(3), pp. 290-293.

10. MENYUK, N., KILLINGER, D.K. and DE FEO, ( ). Appl. Optics 19(19), pp. 3282-3286.

11. MARTHINSSON, B. et al. (1980). Optical and Quantum Electronics 12, pp. 327-334.

12. HESS, R.V. (1981). NASA Langley Research Centre, 25th Annual International Technical Symposium and Exhibit; Physics and Technology of Coherent IR Radar. August 25-26.

13. BROCKMAN, P. et al. (1981). NASA Langley Research Centre, 25th Annual International Technical Symposium and Exhibit; Physics and Technology of Coherent IR Radar. August 25-26.

14. INABA, H. and KOBAYASI, T. (1975). Optical and Quantum Electronics 7, pp. 319-327.

15. HAMZA, M. et al. (1981). Optical and Quantum Electronics 13, pp. 187-192.

16. MENZIES, R.T. and SHUMATE, M.S. (1976). Appl. Optics 15(9), pp. 2080-2084.

17. REMSBERG, E.E. and GARDLEY, L.L. (1978). Appl. Optics 17(4), pp. 624-630.

18. MEGIE, G. and MENZIES, R.T. (1980). Appl. Optics 19(7), pp. 1173-1183.

19. LYNN, D.A. (1976). Air Pollution. Threat and Response. Addison Wesley, New York.

20. PETHERAM, J.C. (1981). 'A comparison between continuously tunable and line tunable CO_2 laser systems for the remote sensing of atmospheric trace gases'. Report prepared at Hull University, Department of Applied Physics, England, for the National Physical Laboratory, Teddington, England.

21. CICERONE, R.J. et al. (1974). Science 185, p. 1165.

22. McCLATCHEY, R.A. and D'AGATI, A.P. (1978). U.S. Air Force Geophysics Labs. Report No. +R-78-0029.

23. PETHERAM, J.C. Private Communication.

24. PENNDORF, R. (1957). J. Opt. Soc. Am. 47, p. 176.

25. VAN DE HULST, H.C. (1957). Light Scattering by Small Particles. John Wiley and Son Inc.,

26. PETHERAM, J.C. (1982). Conf. Proc., 11th International Radar Conference, Wisconsin, U.S.A.

27. MOULTON, P.F. and MOORADIAN, A. (1979). Appl. Phys. Lett. 35(ii), pp. 838-840.

28. GREEN, B.A. (1979). Coherent Laser Radar for Atmospheric Sensing. Department of Applied Physics, Hull University, England.

29. MAYER, A. et al. (1978). Appl. Optics 17(3), pp. 391-392.

30. MENYUK, N., KILLINGER, D.K. and MENYUK, C.R. (1982). Appl. Optics 21(18)

31. GREEN, B.A., LAYFIELD, A.M., RYE, B.J. and THOMAS, E.L. (1980). 'Hybrid Laser Intrapulse Stability', Topical Meeting on Coherent Laser Radar for Atmospheric Sensing, Aspen, Colorado, U.S.A., July 1980.

32. GREEN, B.A., LAYFIELD, A.M., RYE, B.J. and THOMAS, E.L. (1980). 'Coherent Dial Data Acquisition System Design', Topical Meeting on Coherent Laser Radar for Atmospheric Sensing, Aspen, Colorado, U.S.A., July 1980.

33. LAYFIELD, A.M., RYE, B.J., THOMAS, E.L. and VAN DIJK, J.W. (1983). 'A Coherent Laser Radar for Differential Absorption and Meteorological Measurements', Quantum Electronics and Electro-Optics (Ed. P. L. Knight), John Wiley and Son Ltd.

34. PETHERAM, J.C. and VAN DIJK, J.W. (1982). 'A High Pressure Laser for Lidar Measurements of Pollutants in the Atmosphere', 11th International Laser Radar Conference, Wisconsin, U.S.A., June 1982.

35. VAN DIJK, J.W., PETHERAM, J.C. and RYE, B.J. (1983). 'A multi-atmosphere carbon dioxide laser for use in differential absorption lidar (DIAL)', SPIE Conference on Coherent Infrared Radar Systems and Applications II, Arlington, Vancouver, Canada, April 1983.

36. VAN DIJK, J.W., KUSTERS, J.F., LAYFIELD, A.M. and RYE, B.J. ( ). 'The Use of TEA and Multiatmosphere CO_2 Lasers in Active Remote Sensing'. Proc. ESA Workshop on Space Laser Applications and Technology, Les Diablerets, ESA Report SP-2-2, pp. 225-230.

37. KUSTERS, J.F., LAYFIELD, A.M., RYE, B.J. and VAN DIJK, J.W. (1986). 'Broadband CO_2 Laser and Software Filtering in Dial'. Proc. Workshop on DIAL Data Analysis and Control, Virginia Beach, U.S.A.

38. CLIFFORD, S.F., OCHS, G.R. and LAWRENCE, R.S. (1974). J. of Opt. Soc. America 64(2), pp. 148-154.

TABLE 1.1 Pollutants and Atmospheric Gases of interest in remote sensing applications. Laser lines are indicated which have been selected for their suitability in measuring the concentrations of the gases shown. In most cases, suitable reference lines are also included. (From [1] and [29])

Laser	Gas	Recommended Laser Lines					
		Measure			Reference		
		Line	λ (μm)	α_m ($atm^{-1}cm^{-1}$)	Line	λ (μm)	α_m ($atm^{-1}cm^{-1}$)
CO_2	Ammonia	R8	10.333	25.8	R12	10.304	0.06
CO_2	Benzene	P28	9.621	2.0	R20	9.272	0.05
CO	1,3-Butadiene	P13	6.215	2.7	-		
CO_2	1-Butene	P38	10.788	1.3	-		
CO_2	Carbon Monoxide	P20 x 2	4.776	8.0	-		
CO_2	Chloroprene	R18	10.261	9.15	P22	9.569	0.10
CO_2	1,2-Dichloro-ethane	P20	10.591	0.52	R16	10.275	0.01
CO_2	Ethylene	P14	10.533	32.14	P28	10.675	1.27
CO_2	Freon-11	R22	9.261	29.2	P18	9.536	0.10
CO_2	Freon-12	P32	10.719	35.7	P12	10.514	0.08
CO_2	Freon-113	P26	9.604	7.7	-		
CO_2	Ethyl-Mercaptan	R26	10.208	0.56	P20	10.591	0.18
HeNe	Methane	-	3.391	6.0	-		
CO_2	Methyl Alcohol	P34	9.676	8.9	-		
CO_2	Monochloro-ethane	R16	10.275	3.3	P20	10.591	0.06
CO	Nitric Oxide	P11	5.215	6.7	-		
CO	Nitrogen Dioxide	P14	6.229	26.8	-		
CO_2	Ozone	P14	9.505	12.0	P24	9.586	0.60
CO_2	Perchloroethylene	P34	10.742	4.9	R24	10.220	0.1
HeNe	Propane	-	3.391	8.0	-		
CO	Propylene	P9	6.069	0.9	-		
CO_2	Sulphur Dioxide	R26	9.240	0.105	P18	9.536	0.005
CO_2	Trichloro-ethylene	P20	10.591	12.6	R20	10.247	0.04
CO_2	Vinyl-chloride	P22	10.612	8.8	R18	9.282	0.05
CO_2	Water Vapour	R20	10.247	2×10^{-3}	R16	10.275	$< 2 \times 10^{-3}$

TABLE 1.2

Component [% by volume]		Content [ppm]	Component [% by volume]		Content [ppm]
Nitrogen	78.09	780900	Hydrogen	0.00005	0.5
Oxygen	20.94	209400	Methane	0.00015	1.5
Argon	0.93	9300	Nitrogen dioxide	0.0000001	0.001
Carbon dioxide	0.0318	318			
Neon	0.0018	18	Ozone	0.000002	0.02
Helium	0.00052	5.2	Sulphur dioxide	0.00000002	0.0002
Krypton	0.0001	1			
Xenon	0.000008	0.08	Carbon monoxide	0.00001	0.1
Nitrous oxide	0.000025	0.25	Ammonia	0.000001	0.01

Note: The concentrations of some of these gases may differ with time and place, and the data for some are open to question. Single values for concentrations, instead of ranges of concentrations, are given above to indicate order of magnitude, not specific and universally accepted concentrations.

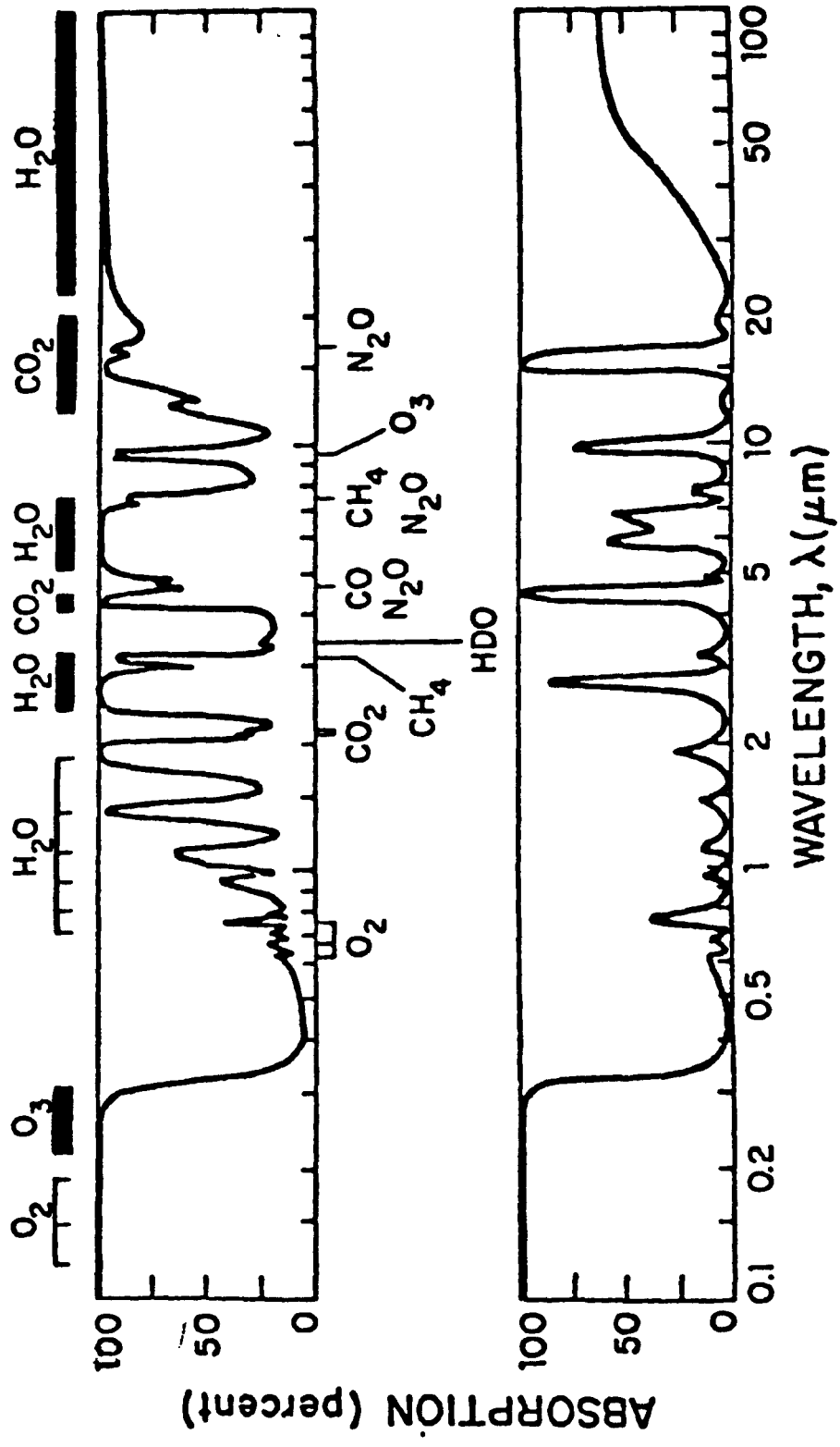
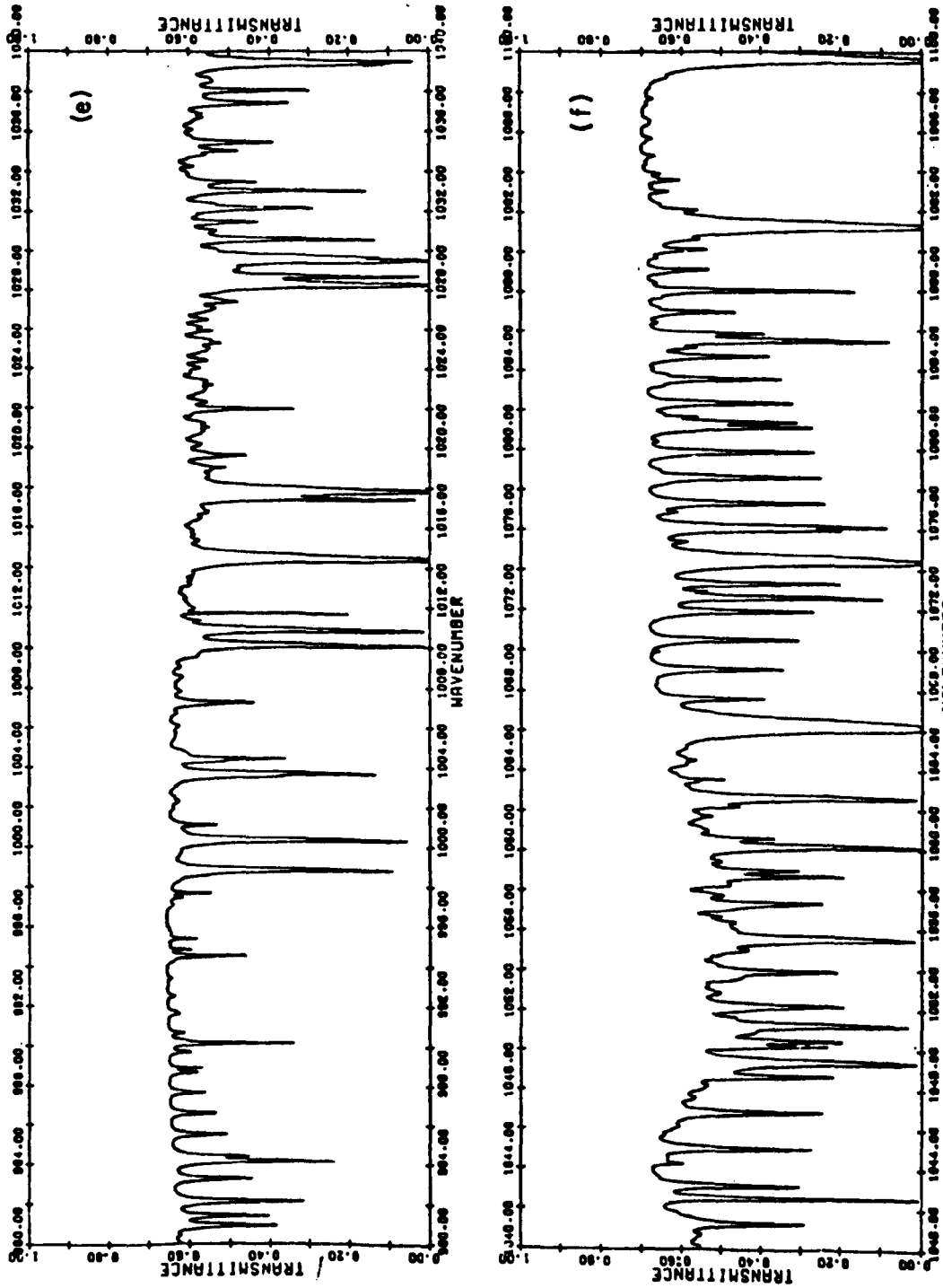
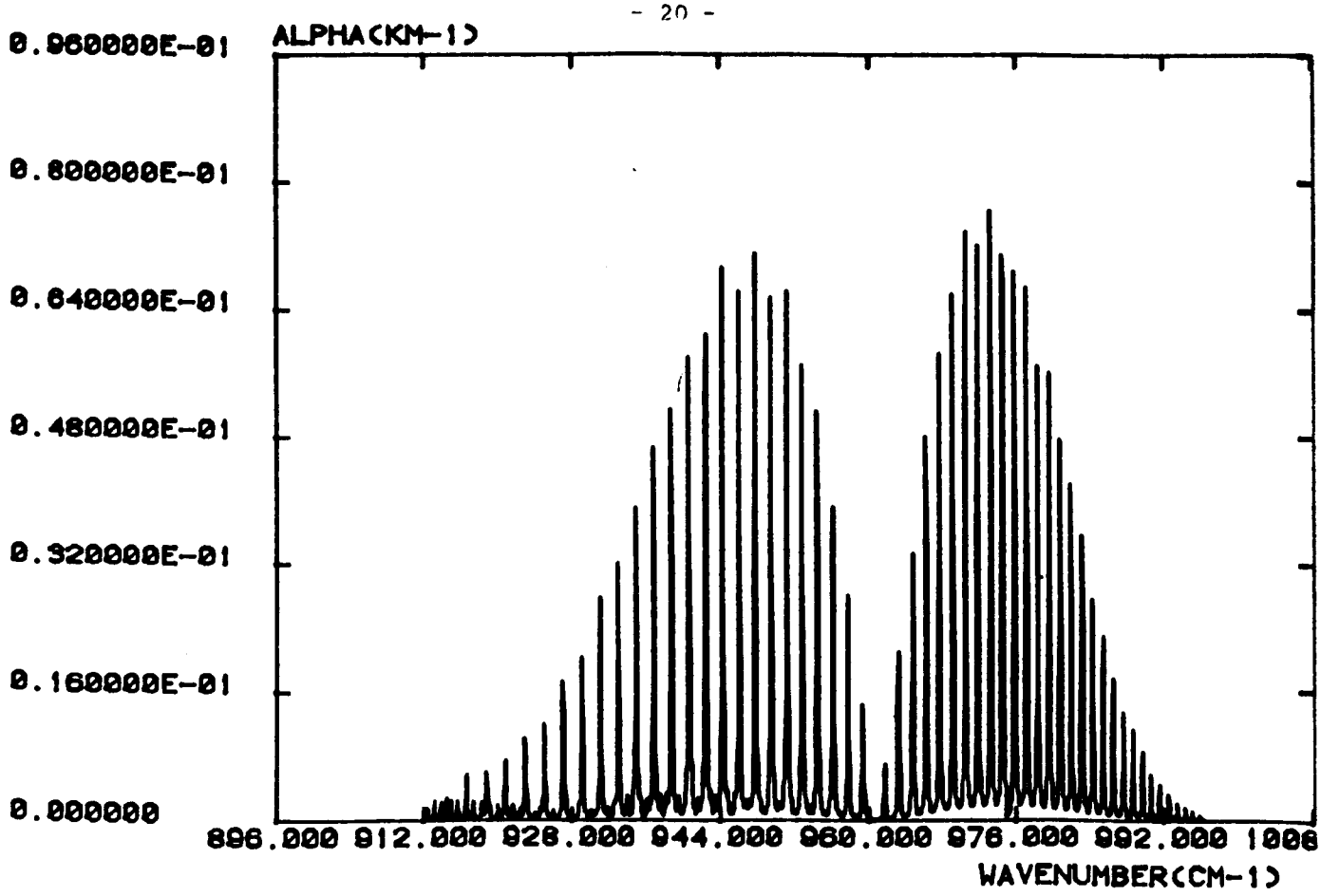


Fig. 1.1 Absorption spectra over a broad wavelength range, both at the Earth's surface and an altitude of 11 km (from [1]).

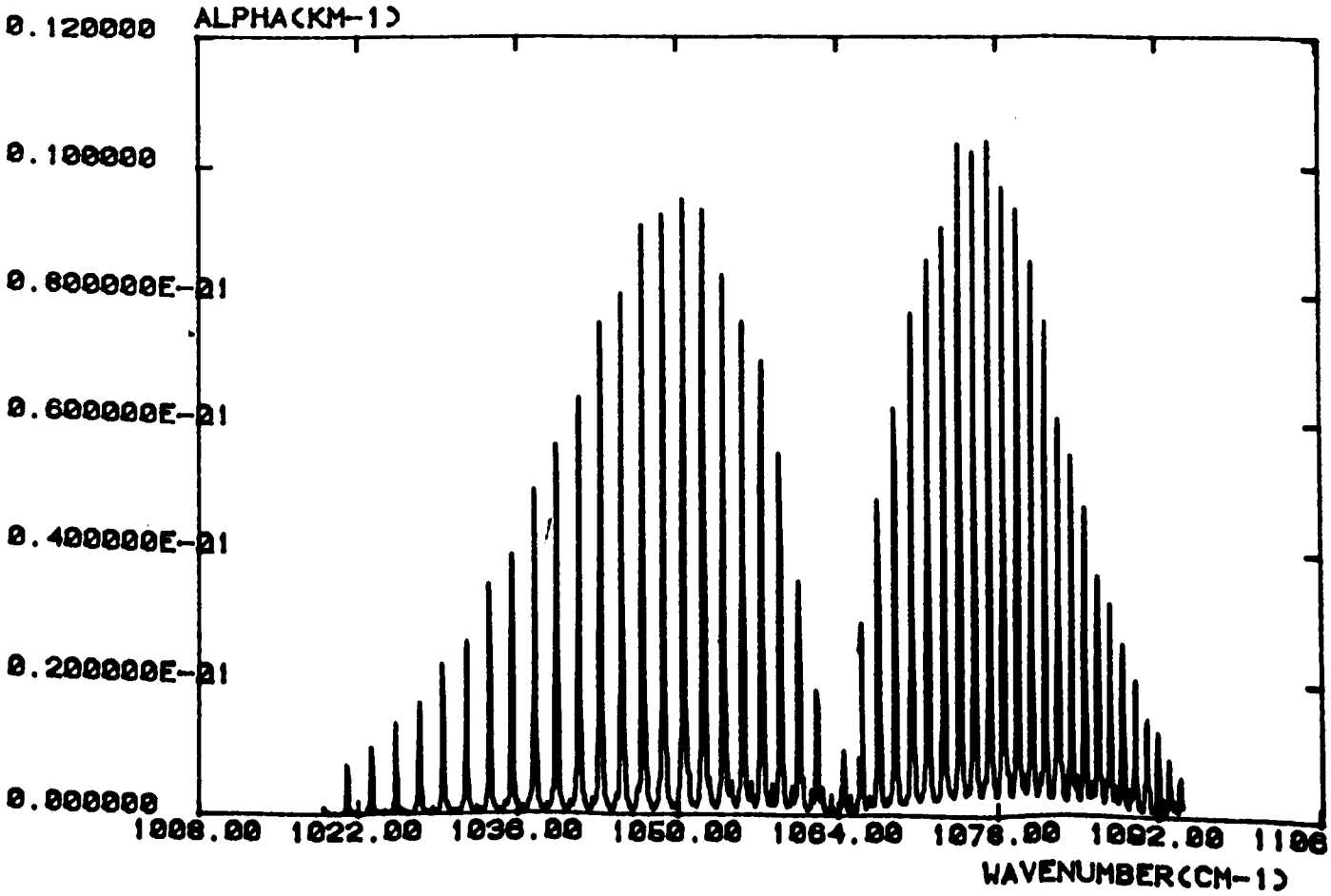


Calculated Spectra for a 10-km Horizontal Sea Level Path. Spectral Resolution is 0.01 cm^{-1}: BOUND value = 20 cm^{-1} (Cont.)

Fig. 1.2



CO2 SPECTRUM 300PPM



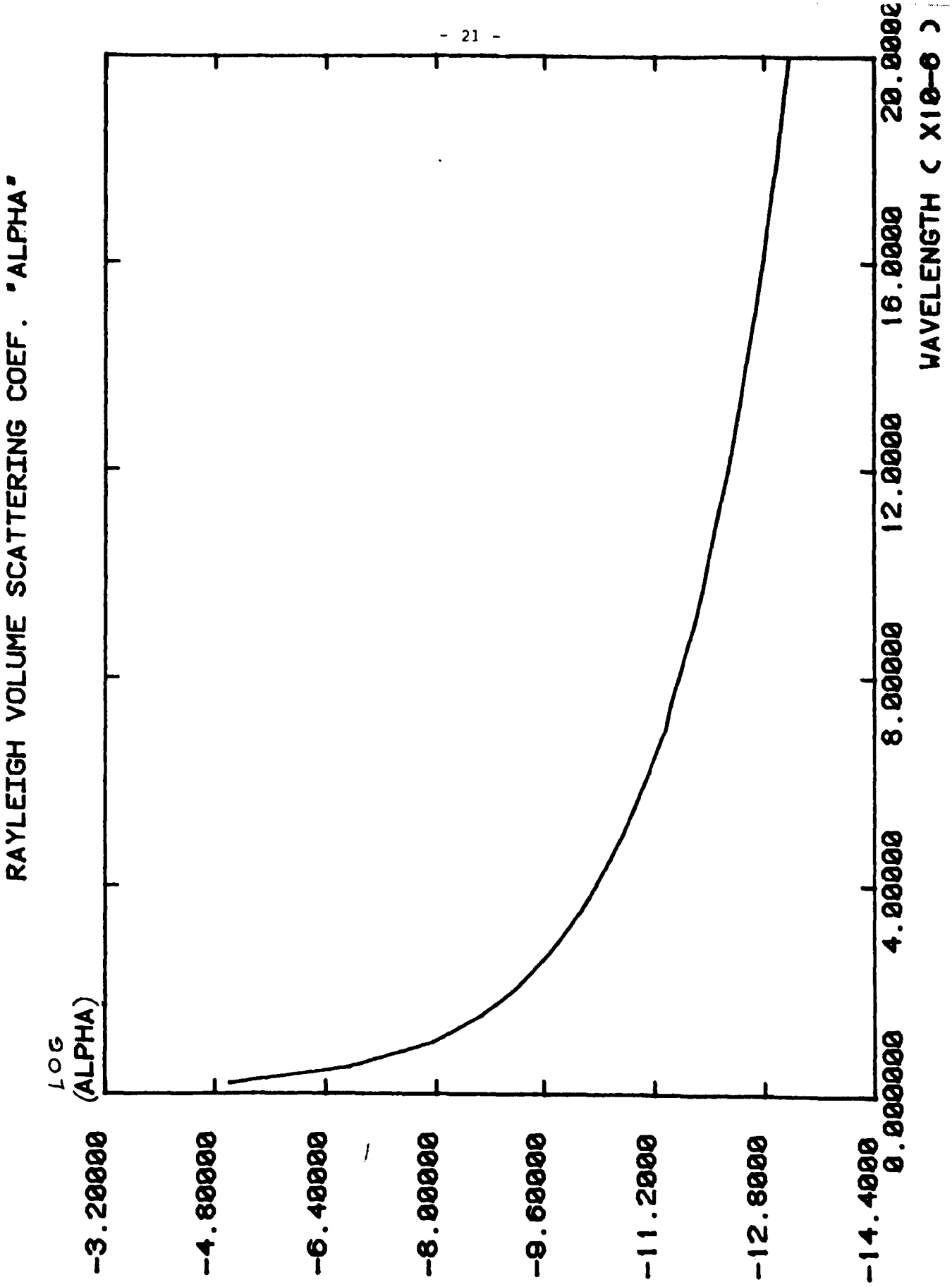

RAYLEIGH VOLUME SCATTERING COEF. 'ALPHA'
LOG (ALPHA)
-3.20000
-4.80000
-6.40000
-8.00000
-9.60000
-11.2000
-12.8000
-14.4000
0.000000
4.00000
8.00000
12.0000
16.0000
20.0000
WAVELENGTH (X10-6)

Fig. 1.4

CHAPTER 2.0

LIDAR SYSTEM HARDWARE

A description of the Hull University lidar system is presented below which gives details of the principal components and the hardware integration necessary for the purposes of the studies reported here. Certain items of equipment, such as the telescopic transceiver and various infra-red detectors were already available. They originally formed part of an earlier lidar system, used to demonstrate heterodyne detection of atmospheric returns. Most of the hardware described, however, was acquired specifically for this project, and most of it was built within the University using the workshops of the Applied Physics Department.

2.1 The Dual CO_2 Laser System

The single, transversely excited atmospheric (TEA) pressure CO_2 laser which acted as the transmitted energy source for the heterodyne lidar work was considered inadequate for direct detection studies due to its low output energies, low pulse repetition frequency (prf) and unreliable power supply. It was apparent at the outset that two alternatives existed, in principal, for generating dual laser pulses, of differing line tunable wavelengths, and at intra pulse pair separations ranging from near simultaneous down to the order of a second:

(1) The use of a single gain section, and the application of wavelength switching, accomplished via the mechanical displacement of a line tuning device (a diffraction grating for example), or

(2) Dual gain sections, and dual laser resonator optics, forming two complete laser systems.

The first option was discarded primarily because of the lower limit to the pulse pair temporal separation achievable with a mechanical switching arrangement. Delays of no less than a few milliseconds probably represents the best that could be achieved with such devices. Since external sources had already indicated an atmospheric decorrelation time of this same order [1], it was considered desirable to be able to produce intra pulse delays of at least an order of magnitude smaller to preserve the correlation in temporally adjacent measurement channels. The significance of retaining high cross channel correlations will be discussed in Chapter 5.

To simplify the task of generating near identical outputs in all respects other than wavelength, and with pulse pair separations down to a few tens of microseconds, it was decided that two new TEA gain sections should be constructed, both having identical specifications. An acknowledged disadvantage of the dual laser system is the differing propagation path lengths that are automatically introduced between the two laser outputs and the optics of the transmitter telescope. Other considerations which influenced this decision, however, derive from the potential flexibility of a system with dual gain sections.

Both TEA gain sections could, in principle, be operated simultaneously and configured into a single laser system, or else as a laser-amplifier arrangement. Such flexibility made feasible several alternative systems:

(i) A single laser lidar system of increased transmitted energy.

(ii) A long cavity laser system for increased frequency diversity in the longitudinal modes; this application is specific to

heterodyne detection and would have the statistical advantage that each mode may be optically mixed with a local oscillator to produce as many independent samples as there are modes.

(iii) A short pulse laser system. Optical switching would be used to extract a pulse section of short duration from the peak output of a laser based on one of the gain sections. This would subsequently be amplified by the remaining gain section. Short pulse techniques represent an alternative approach to the problem of increasing the number of independent samples collected per pulse pair. This technique, however, is specific to atmospheric backscatter returns since the extra samples would derive from the relative spatial extent of the effective laser pulse length (see Chapter 3), and the required range resolution for a given precision in the gas concentration estimate.

Although the last two are of relevance principally to lidar systems employing heterodyne detection, early experiments were conducted to examine the feasibility of these techniques. However, since the primary objective of these lidar system investigations was gas concentration measurement, these alternatives were passed over in favour of a system more directly applicable to this requirement. They remain, therefore, potentially exploitable methods for improving the number of independent samples collected, per pulse pair, particularly for range resolved lidar work.

2.1.1 The TEA Discharge Gain Sections

A schematic of the basic capacitor charging system used for

each gain section is provided in Figure 2.1(a). Although an RC circuit is not the most efficient method of capacitor charging, compared to ready built, commercially available constant power systems, it was relatively inexpensive. The power supply used was purchased from Universal Voltronics and was rated at 50 kV and 375 mA, thus providing a maximum dc output power of 18.75 kW. Illustrated in Figure 2.1(b) is a cross section of the TEA gain sections showing the aluminium electrodes (shaded), and the gas circulating fans enclosed in a glass barrel which is sealed at either end using 1" thick perspex discs. Dimensions are as indicated. Not illustrated is the ultra-violet pre-ionisation system (discussed later), and the copper water cooling tubes which are located directly behind the fans.

The circuit element at the right hand extremity of Figure 2.1(a) is intended to depict both a physical and electrical cross-section of the TEA discharge region. The primary discharge plasma is confined to the space between the two opposing "D" shaped electrodes. These are separated by a distance of 2.5 cm and are intended to define a uniform electrical field over a lateral extent of 3 cm. Figure 2.1(b) indicates that the active length of the electrodes is 50 cm.

A pre-ionisation discharge is initiated at some position remote from the main discharge volume (approximately 10 cm); this is illustrated in Figure 2.1(a) as a pair of opposing pins, positioned either side of the main electrodes, and connected to the high voltage power supply via 270 pF capacitors. Each gain section possessed eleven pairs of UV pre-ionisation pins, which were distributed along the length of the electrode with five on one side and six on the other (staggered rather than opposing each other). The pre-ioniser discharge serves as a source of strong UV photon radiation which propagates into the main

discharge volume and pre-ionises a fixed number of neutral laser gas molecules in this region. Electrons resulting from this pre-ionisation serve to condition, both spatially and temporally, the main discharge plasma.

In operation, the 0.05 μF energy storage capacitor C_e is charged to a voltage, V_c, close to the power supply voltages, V_o. A trigger voltage is then supplied to the spark gap, which acts as a switch, thus completing the circuit which includes the capacitor C_e, pre-ioniser capacitors C_p, the main discharge electrodes, a 50 ohm pulse shaping resistor, and an inductance, L, representative of the electrical connections between the gain section and the charging circuit. Figure 2.2 illustrates alternative schematic views of the relative physical locations of the various components identified in Figure 2.1. By selecting appropriate values for the UV pre-ioniser capacitors, C_p, and by adjusting the pre-ioniser pin separations, it is possible to arrange that the main discharge is initiated after some fixed time delay, τ_d, with respect to pre-ioniser initiation.

τ_d must, of course, fall within the rise time of the voltage across the main electrodes which was observed to be approximately 500 nsec. Apart from this restriction, however, discharge uniformity seemed relatively insensitive to the selected value of τ_d. At low prfs it also appeared to be relatively immune to variations of approximately 50% in the number of UV pre-ioniser pins used. For the 11 pin configuration described here, 6% of the total discharge energy was used by the UV pre-ionisers. This percentage is somewhat higher than that reported for other UV pre-ionised laser systems [$\sim$ 1% for ref. 2] but was retained in the lasers described here as a precaution against possible arcing during lengthy periods when the lasers were to be fired at an

uninterrupted prf of 10 Hz.

An analysis of the circuit of Figure 2.1(a) reveals that the instantaneous current into the main discharge plasma is given by [2]

$$i(t) = (V_c - V_p)\sqrt{\frac{C_e}{L}}\ \sin w_o t \qquad (2.1)$$

where $w_o = (LC)^{-\frac{1}{2}}$ and V_p is the plasma discharge sustaining voltage. V_p is a function of the electrode separation, d, gas pressure, p, and gas mixture. In general, the value of V_p/pd or E/p should be constant for quasi-steady-state operation of the plasma in the avalanche mode [2].

The peak value of the current and the time to reach maximum value are given by

$$i_{max} = (V_c - V_p)\sqrt{\frac{C_e}{L}} \qquad (2.2)$$

and

$$t_{max} = \frac{\pi}{2}\sqrt{LC} \qquad (2.3)$$

t_{max} was observed to be approximately 500 nsec for each of the gain sections, so the induction L is equal to 2μH, and the resistance term $\sqrt{(C_e/L)}$ has an approximate value of 0.16 ohms. The energy input to the plasma, E_p, may be obtained by multiplying (2.1) by V_p and integrating over a time interval equivalent to twice the current rise time ($2 \times t_{max}$). When $w_o t = \pi$, the value of E_p becomes [2]

$$E_p(\pi) = 2\, C_e(V_c - V_p)V_p \qquad (2.4)$$

and the energy transfer efficiency η_p, defined as the ratio of E_p to the total energy stored in the capacitor, C_e ($= \frac{1}{2} C_e V_c^2$) is then

$$\eta_p = 4\,\frac{V_p}{V_c}\left(1 - \frac{V_p}{V_c}\right) \qquad (2.5)$$

This is plotted as a function of V_p/V_c in Figure 2.3 (after Judd and Wada, 1974).

Complete transfer of energy into the plasma is obtained only if $V_p/V_c = 0.5$, which is equivalent to stating that impedance of the LC network must be matched to that of the plasma. If it is assumed that the voltage across the energy storage capacitor C_e is the same as the power supply voltage, V_o (= 40 kv), then the maximum energy that can be transferred to the plasma will be, from (2.4), $E_p(max) = 40$ Joules. This energy is deposited into a plasma volume of 3 cm x 2.5 cm x 50 cm = 0.375 L, giving an energy density of $\sim$ 100 Joules/L. Electrical to optical energy conversion efficiency is dependent on the ratio of the optical mode volume to the discharge plasma volume but if it is assumed that, in the absence of any mode control, 10% of the plasma discharge energy is extracted as optical energy (a typical value for pulsed CO_2 laser systems) then the maximum output of the laser will be $\sim$ 4 Joules.

The 100 k ohm resistors of the RC charging circuit were selected on the basis of the required prf from each gain section and the output rating of the power supply. Since the instantaneous voltage across the energy storage capacitor is given by

$$V_c(t) = V_o(1 - e^{-t/RC_e}) \tag{2.6}$$

the time for V_c to reach a given fraction of V_o becomes

$$t = - RC_e \ln \left[1 - \frac{V_c(t)}{V_o} \right] \tag{2.7}$$

If V_c is to achieve 99% of V_o, then t = 0.02 seconds, imposing an upper limit of 50 Hz on the prf per gain section. At this frequency, the mean current drawn from the power supply will be

$$i = \frac{dQ}{dt} = prf.V_cC_e = 100 \text{ mA} \tag{2.8}$$

and the power dissipated in the charging resistors is then

$$P_R = i^2 R = 1 \text{ kW} \tag{2.9}$$

These values will, of course, be doubled for dual gain section operations, but are still well within the dc rating of the power supply.

All of the charging components illustrated in Figure 2.1(a) were replicated for each gain section and the entire assembly immersed in an oil filled PVC tank measuring 42" x 24" x 24", situated underneath the gain sections (see Figure 2.2). In addition to providing electrical insulation, this large reservoir of oil provided a means of heat dissipation for the resistor stacks, but at the 10 Hz prf employed for the lidar measurements, this was not a severe problem.

2.1.2 Laser Design

Practical guidance as to the selection of suitable design parameters was available via, for example, the published work of Andrews, Dyer and James [2] in which a simplified set of rate equations are developed, describing how resonator mirror reflectives and the molecular gas mixture ratio influences the laser power and energy. They obtain, as the useful laser power per unit mode volume,

$$P_o = - h\nu q \frac{cT \ln R}{2L(T+A)} \tag{2.10}$$

where $h\nu$ is the photon energy of the laser transition, q is the photon density in the laser cavity, c is the speed of light, L is the cavity length, and R, T and A are the output mirror reflection, transmission and absorption coefficients respectively. In deriving (2.10), the growth in the photon density is averaged over the entire laser cavity. Also, no account is taken of laser mode properties such as spatial variation in

the photon density transverse to the cavity axis, or axial mode beating [2].

Laser output energy may be determined by integrating P_o over the duration of the pulse τ_p;

$$E_p = \int_o^{\tau_p} P_o \, dt \tag{2.11}$$

CO_2 laser pulses consist of a high amplitude gain switched spike followed by a relatively low amplitude tail extending for several microseconds. Figure 2.4 illustrates a digitised record for the reference transmitted power, and the power received from a topographic target, monitored using an LTT detector. Details of the experimental arrangement used to obtain such figures will be discussed later in this and subsequent chapters but is referred to now because of the pulse shape information it contains. Sample numbers along the abscissa are multiples of 10 nsec so that the total pulse duration, including tail, is 4 μsec. The gain switched spike, at "full width half maximum " (FWHM), however, has a duration of approximately 350 nsec. Therefore, if P_o (max) corresponds to the peak power of the gain switched spike and τ_p is its full width at half maximum, then the energy contained in the spike may be approximated by forming the product,

$$E_p = P_o(\max)\ \tau_p \tag{2.12}$$

Gas composition is characterised by the molecular gas ratio

$$\gamma = \frac{w}{(x+w)} \tag{2.13}$$

where w is the CO_2 partial pressure and x is the N_2 partial pressure. A third gas, helium, tends to form the dominant component of the gas mix but, unlike CO_2 and N_2, it is not directly involved in the energy transfer

processes. Its function is to depopulate one of the lower energy levels of CO_2 to which non-radiative transitions occur.

Assuming a gain length of 70 cm and a cavity length of 120 cm, Andrews et al obtain results for two cases which may be summarised as;

(a)	(b)
$\gamma = 0.5$	$\gamma = 0.1$
$R = 0.7$	$R = 0.9$
$R' = 0.975$	$R' = 0.975$
$A = 0.01$	$A = 0.01$
$P_o(max) = 11\ kW\ cm^{-3}$	$P_o(max) = 2kW\ cm^{-3}$

The molecular gas ratio used in the dual gain section of the lidar system was $\gamma = 0.33$ and both output mirrors had a reflectivity of $R = 0.8$. Since these values lie midway between those pertaining in cases (a) and (b), an interpolated value for the photon density, q, may be obtained by substituting the listed parameters for (a) and (b) into equation (2.10) and solving for q. If this is done it is found that

$$q_a = 1.276 \times 10^{16}$$

and

$$q_b = 8.433 \times 10^{15}$$

so that $q = \sqrt{q_a q_b} \simeq 1.0 \times 10^{16}$

The value of $P_o(max)$ derived using this value of q in (2.10) is then $P_o(max) = 5.3\ kW/cm^3$.

An estimate for the integrated laser output power is now dependent on the dimension of the beam mode volume. Using the generally applicable

beam propagation equations presented later in this chapter, a computer generated estimate, illustrated in Figure 5.6(a), reveals a relatively constant intra-cavity beam diameter of approximately 1 cm for the TEMoo Mode (R_1 = 0.8, R_2 = 1.0 (a diffraction grating) and radius of curvature for R of 20 metres). Therefore, assuming that the beam mode volume may be approximated by a cylinder of radius 5 mm and length 2.4 metres, the peak pulse output power becomes

$$P_{max} = \pi r^2 L P_o(max) \simeq 1 \text{ MW}$$

giving a gain switched pulse energy of 0.35 Joules, using (2.12). In practice the actual energies measured for each gain section, at low prfs, tended to fall in the range 100 mJ to 150 mJ but, unlike the cases considered by Andrews et al, these values were obtained for cavities in which R' is a diffraction grating. The losses associated with this component, which was included to provide a line tuning capability, will be significantly higher than for a mirror. Furthermore, at sustained prfs of 10 Hz, the output power was observed to decrease by as much as 50% over experimental runs totalling 10,000 shots.

It will be shown in Chapters 3 and 4 that, for topographic target returns, transmitted pulse power is more important than transmitted pulse energy. This is in direct contrast to the situation for distributed (atmospheric) targets where the pulse energy is more significant. Variations in laser energy and peak power as a function of the molecular gas mixture ratio γ are illustrated in Figure 2.5 (after Andrews, Dyer and James [3]) for various output mirror reflectivities. Laser energy is denoted by the broken lines and laser power by full lines. It is apparent that a molecular gas ratio of 0.33 optimises the lasers for energy output rather than for power.

Other factors, however, became important in the selection of γ.

Foremost was the inhibition of arcing between the laser electrodes during experiments consisting of 10,000 shots. It was found that lower values of γ enabled approximately twice as much data to be collected before arcing would commence. Frequent arcing was found to damage the electrodes since it tended to persist in those regions where it had started. This would necessitate a complete strip down of the resonator and gain section to permit re-polishing of the electrode surfaces; a very time-consuming operation which it became paramount to avoid. Higher energy outputs were also found to confer advantages during alignment of the system, since the response of carbon coated beam locators appeared to be energy dependent.

Hybridisation

Hybridisation refers to the technique of incorporating a low pressure CW CO_2 gain section in the same resonator as the TEA gain section. As far as a direct detection lidar is concerned, its principal advantages are;

(1) Amplification is forced to occur at those frequencies near to the centre of the atmospheric pressure gain curve.

(2) Selection of a single axial mode.

Figure 2.6 is a schematic cross-section of each hybridised laser showing both gain sections, their electrodes (cross hatched) the sodium chloride Brewster windows used to seal both gain sections, the diffraction grating and the zinc selenide output coupler. The low pressure section has an active length of approximately 80 cm.

A TEA CO_2 laser without hybridisation has an emission line width of approximately 3 GHz (at FWHM) but a low pressure gain section, operated

at 15 - 20 mbar has a line width of approximately 100 MHz (FWHM). Separation between adjacent lines is some 30 to 60 GHz, depending on wavelength. The laser resonator itself, however, possesses a characteristic axial mode spacing which is dependent on the cavity length;

$$\Delta\nu_{axial} = \frac{c}{2L} \qquad (2.14)$$

For L = 2.4 metres, $\Delta\nu_{axial}$ = 60 MHz. Therefore, over a 3 GHz line width, lasing action may potentially occur over approximately 50 axial modes. This may be contrasted with the situation for the low pressure gain section where only one axial mode will be amplified sufficiently to overcome various loss mechanisms. By preferentially amplifying a single axial mode, therefore, all of the energy derived from the pulse gain section is confined to within $\sim$ 100 MHz, thus guaranteeing the spectral purity of the lidar system probe wavelength.

Figure 2.7(a) presents a plan view of the dual laser system with both sets of gain sections installed within a single resonator. The resonator was constructed using lengths of Oriel optical railing, the bores of which were filled with water to maximise the thermal inertia and to attenuate acoustic vibrations. Invar rods, 40 mm in diameter and over 2 m long, were inserted inside the longest sections of this rail as a further aid to thermal stabilisation. These were clamped to the transverse rails at both ends of the resonator, but were permitted to expand through the water tight seals at the ends of the longest rails via a series of "O" rings. Attached to the front of the resonator are extra sections of railing used to carry the components which couple both laser outputs into a single optical path.

Optical components depicted in Figure 2.7(b) include, at one extremity of the resonator, the grating mounts and a pair of adjustable

irises used for transverse mode control. At the other extremity are a complementary set of irises, the output couplers, a mirror, a beam splitter, and two beam aligners used for both translational and rotational adjustments to the beam path. Figure 2.8 is a photograph taken at an early stage in the testing of the dual laser system showing the TEA and low pressure gain sections, mounted above the oil filled tank containing the charging circuits. Visible in the foreground are two 40 mm diameter invar rods used to stabilise the resonator.

2.1.3 Resonator Configuration and Beam Propagation [4,5,6,7]

The properties of laser resonators tend to be characterised in terms of so-called "g" parameters which are defined to be

$$g_1 = 1 - \frac{L}{R_1} \quad \text{and} \quad g_2 = 1 - \frac{L}{R_2} \tag{2.15}$$

where R_1 corresponds to the radius of curvature of the output mirror (coupler) and R_2 to the radius of curvature of the opposite mirror which ideally, has a reflectivity of unity. R_1 and R_2 are defined to be positive if the mirrors are concave with respect to the resonator interior. A stability condition exists for resonators which is expressed as

$$0 < g_1 g_2 < 1 \tag{2.16}$$

Mirror combinations which lie outside this region tend to originate diverging wavefronts which effectively focus all energy away from the mirror at the opposite end, resulting in high diffraction losses.

For the dual resonator considered here, $g_1 = 0.88$ and $g_2 = 1.0$, which is tending towards one extremity of the stability limit. This is a consequence of the large radius of curvature (ROC) of the output mirror

(R_1 = 20 m). Long radius mirror configurations have the advantage of supporting propagation modes of almost constant beam diameters which can be matched to the "bore" dimensions of the discharge volume, thus maximising the energy extraction efficiency. They also produce beam diameters, exterior to the cavity, which exhibit relatively little divergence over propagation paths of at least several metres. Figure 5.6(b) reveals that the computed beam diameter for the 20 m ROC cavity increases by only $\sim$ 50% over a path length of $\sim$ 5 metres. Such constancy is useful both in minimising the number of refractive components required between laser output and telescopic transmitter, and for maintaining some parity in the beam propagation losses between outputs from two lasers separated by approximately 1 m.

Although diffraction losses increase as the stability limit is approached, for a 20 m radius of curvature, these losses were found to be small compared to those attributable to other factors (such as the salt flats sealing both high and low pressure gain sections). A parameter used to characterise these diffraction losses is the "Freznal Number", N, which for a circular mirror of radius, a, is given by

$$N = \frac{a^2}{L\lambda} \qquad (2.17)$$

where λ is the laser wavelength. Since an aperture of approximately 15 mm was imposed by the transverse mode control irises, the value of N is found to be 2.34. Diffraction losses vary inversely with the magnitude of N and are therefore minimised for large values.

Mode Spacing

The spatial distribution of energy both inside and outside (near field) a resonator is described by its TEMmnq mode number where m and n

refer to transverse modes, and q is the axial mode index. m, n and q are measures of the number of half wavelengths along each mode axis. In general, the resonance frequency for any mode combination, mnq, is given by

$$\nu_{mnq} = \left[q + (m+n+1) \frac{\mathrm{Cos}^{-1}\sqrt{g_1 g_2}}{\pi} \right] \frac{c}{2nL} \qquad (2.18)$$

where n is the refractive index of the lasing medium. The mirror combination used for each of the lasers falls between two well defined resonator types; the "plane parallel" and the "hemispherical" resonators. The former is perhaps more appropriate since the latter category applies where the mirror spacing, L, is slightly less than R_1. However, for the plane resonator (2.18) reduces to

$$\nu_{mnq} = \frac{q_c}{2nL} \qquad (2.19)$$

and for the hemispherical resonator,

$$\nu_{mnq} = \left[q + \frac{(m+n+1)}{2} \right] \frac{c}{2nL} \qquad (2.20)$$

If mode spacings are now considered, the separation between adjacent axial modes, for both resonator types, becomes

$$\nu_{mn\Delta q} = \Delta\nu_{mn(q+1)} - \Delta\nu_{mnq} = \frac{c}{2nL} \qquad (2.21)$$

and between adjacent transverse modes is

$$\nu_{\Delta mnq} = \left(\mathrm{Cos}^{-1}\sqrt{g_1 g_2}\right) \frac{c}{2\pi nL} \qquad (2.22)$$

Expression (2.21) has already been used in the justification of hybridisation but it is of general interest to note here that the transverse

mode spacing for the dual laser system is approximately 7 MHz. This may be compared with the 60 MHz separation between axial modes.

Finally, the bandwidth associated with each resonant mode ν_{mnq} is

$$\Delta\nu_{mnq} = \frac{}{2\pi\tau_c} = \frac{c(\alpha L - \ln\sqrt{R})}{2\pi n L} \qquad (2.23)$$

where τ_c is the photon lifetime within the cavity, R is the reflectivity of the output mirror, and αL is the fractional loss in beam intensity per cavity pass. A precise value for αL was not determined for the lasers considered here but if it is assumed that αL has a maxmimum value of 0.1, then the bandwidth becomes approximately 4 MHz. Consequently, for the TEMooq mode, the output laser energy at any instant will actually be confined to within this bandwidth but, between laser shots, the centre frequency will drift over the $\sim$100 MHz frequency interval defined by the low pressure CW section.

Gaussian Beam Propagation

The role of the mode control irises, mentioned previously was to confine the laser output, for the purposes of these experiments, to the TEMooq mode in which the variations in amplitude across the beam wavefront are described by the Gaussian distribution function. This mode has the principal advantage of minimising beam angular divergence. Briefly presented below are the various equations which define important Gaussian beam parameters, both inside and outside the laser resonator, and which were used in a computer program, LASER1.FOR, to predict beam propagation characteristics throughout the lidar system.

The mean radius of the TEMooq mode is variously described as the beam radius or "spot size" but is generally denoted as w. It is defined

to correspond to the distance from the mode centre at which the electric field amplitude has fallen to 1/e of its mode centre value (e is the Naperian logarithmic base). The relationship of w to the physical parameters of the resonator was first derived for "confocal" resonators by Boyd and Gordon [8] and then generalised to other resonators by Boyd and Kogelnik [9].

Beam diameters reach a minimum value known as the "beam waist", w_o, which is characteristic of the optical components between which the beam is propagating. Inside a resonator this is given by

$$w_o = \left(\frac{\lambda L}{\pi}\right)^{\frac{1}{2}} \frac{[g_1 g_2 (1 - g_1 g_2)]^{\frac{1}{4}}}{(g_1 + g_2 - 2g_1 g_2)^{\frac{1}{2}}} \tag{2.24}$$

and it is located at a distance

$$z_1 = \frac{-g_2(1 - g_1)L}{g_1 + g_2 - 2g_1 g_2} \tag{2.25}$$

from the non-transmitting mirror. The spot size at this mirror is given by

$$w_1 = \left[\frac{\lambda L}{\pi}\right]^{\frac{1}{2}} \left[\frac{g_2}{g_1(1 - g_1 g_2)}\right]^{\frac{1}{4}} \tag{2.26}$$

but at the output mirror it becomes

$$w_2 = \left[\frac{\lambda L}{\pi}\right]^{\frac{1}{2}} \left[\frac{g_1}{g_2(1 - g_1 g_2)}\right]^{\frac{1}{4}} \tag{2.27}$$

Outside the resonator, the beam radius at a distance, z, from the beam waist is

$$w(z) = w_o \sqrt{1 + \left[\frac{\lambda z}{\pi w_o^2}\right]^2} = w_o \sqrt{1 + \left(\frac{z}{z_R}\right)^2} \tag{2.28}$$

where

$$z_R = \frac{\pi w_o^2}{\lambda} \tag{2.29}$$

is referred to as the "Rayleigh Range". Another property of Gaussian beams is their spherical wavefronts, which have radii dependent on the propagation distance, z. This behaviour is described by

$$R(z) = z\left[1 + \left[\frac{\pi w_o^2}{\lambda z}\right]^2\right] = z\left[1 + \left[\frac{z_R}{z}\right]^2\right] \tag{2.30}$$

The Rayleigh Range, z_R, is used to define two domains for the approximate solution of many optical propagation calculations. If z is the distance from the resonator beam waist such that

$$z >> z_R \tag{2.31}$$

then z is said to be in the "Far Field", but for values of z in the region

$$z < z_R \tag{2.32}$$

the term "Near Field" is used. An immediate consequence of this definition is that the far field spot size, ordinarily given by (2.28), reduces to

$$w_{FF}(z) = \frac{z\lambda}{\pi w_o} \tag{2.33}$$

Also, the half angle beam divergence (θ), in the far field, becomes

$$\theta = \frac{\lambda}{\pi w_o} \tag{2.34}$$

Using expression (2.24) the beam waist calculated for the lasers considered here is approximately 4.5 mm. As a result, the Rayleigh Range is found to be 6.36 m and the beam divergence, θ, is 0.7 mrads.

At the telescopic transmitter, focusing components are encountered of both the reflective and refractive type which may be approximated by positive thin lens formulae. The beam waist location for a focused laser beam is therefore

$$z_2 = f + \frac{(z_1 - f)f^2}{\left[(z_1 - f)^2 + \left(\frac{\pi w_{o1}^2}{\lambda}\right)^2\right]} \tag{2.35}$$

where f is the focal length of the "lens" and w_{o1} becomes the beam waist prior to this focusing component. If w_{o2} is used to denote the beam waist after the propagating beam has encountered the lens, then

$$\frac{1}{w_{o2}^2} = \frac{1}{w_{o1}^2}\left(1 - \frac{z_1}{f}\right)^2 + \frac{1}{f^2}\left(\frac{\pi w_{o1}}{\lambda}\right)^2 \tag{2.36}$$

where z_1 and z_2 are the separations between the lens and the beam waists, w_{o1} and w_{o2} respectively.

Lastly, of significance in determining energy losses throughout the lidar system, is an expression which relates the power P(a) transmitted through a circular aperture of radius, a, to the power, P_i, incident on the aperture; this is

$$P(a) = P_i[1 - \exp(-2a^2/w^2(z))] \tag{2.37}$$

where w(z) is the beam radius at the aperture. Figures 5.6(a) through to 5.6(d) present the results of applying the above relationships to the propagation paths occurring in the lidar system. Beam diameters are plotted as a function of propagation distance out to the topographic target at 1.8 km. Energy losses are included at the locations of the appropriate aperturing components. These figures will be considered further in Chapters 4 and 5.

2.2 Optical System Integration

Figure 2.9 provides a plan view of the lidar system revealing the relative locations of the dual hybridised laser system and the optical transceiver in two adjacent laboratories. A stair well separates the two rooms but a 6" diameter plastic pipe was installed to admit the laser output into the "telescope room". Not illustrated is a third, adjacent laboratory, which housed the computer, some other signal processing electronics and the front end unit of a laser trigger system. The co-axial telescope is shown offset slightly with respect to the laboratory to align its optical axis with a convenient topographic target.

The optical propagation path may now be described with reference to Figures 2.9 and 2.7. After passing the laser output coupler each propagating wavefront encounters a beam aligner which is used to provide both translational and rotational adjustments to the beam path. These devices were a copy of a commercially available unit for aligning the outputs of HeNe lasers, but the dimensions were doubled in the versions described here to avoid beam truncation. The output from laser 2 is reflected along an optical rail and passes through a 5 cm diameter beam splitter (germanium) to the first of a pair of irises used to define the beam path once the system has been aligned. The output from laser 1 is then reflected by the beam splitter and is also directed towards this same iris (labelled Iris 1 in Figure 2.7).

Two alternatives were available for coupling the outputs from both lasers into a single optical path. The first, and most energy efficient, method would have been to install an optical flat adjusted to the Brewster angle and rotate the plane of polarisation of one of the lasers so that each laser satisfied one of the polarisation requirements for transmittance or reflectance by the Brewster plate. After some initial trials

with a germanium plate 75 mm in diameter and 5 mm thick, it became apparent that this approach would introduce significant extra complexity in the alignment procedure. A 50% beam splitter was therefore selected instead, even though 50% of the beam energy is lost, because of the advantages it conferred in terms of alignment. The optical component removed from the beam path was, however, used to monitor the output power from the low pressure gain sections to ensure that CW laser action did not fall below 50% of its peak value (∿1 watt).

After passing through the connecting tube, a 100% mirror is encountered which directs the outputs from both lasers through the second beam aligning iris. These two components are located in Figures 2.7(b) and 2.7(c). The first figure presents a side view of the primary optical rail used in the telescope room to carry all of the optical components required for direct detection. Figure 2.7(c) is a plan view of the same rail (below the telescope) coupled together with two shorter, parallel rails which were planned to carry CW local oscillators for future heterodyne detection studies.

Beyond this second iris, the beams continue through a hole bored in the vertical section of optical rail, and are then reflected upwards by a 90% beam splitter towards the "transmitter (Tx) lens" which is used to couple the laser energy into the transmitting telescope. A second germanium beam splitter is used here, rather than a 100% mirror, to permit 10% of the transmitted energy to be monitored by a mercury cadmium telluride (CMT) detector located further along the optical rail. This site was selected for monitoring the transmitted power, in preference to the more obvious site at the 50% beam splitter used in the laser room, to minimise the electrical noise interference generated by the laser trigger and discharge pulses.

The Tx lens located above the second beam splitter was made of zinc selenide and had a focal length of 100 mm. Its transparency to both visible and I.R. radiation was a useful aid during the initial alignment stage since HeNe lasers could be used. In coupling the output power into the transmitter telescope, its primary function was to cause both beam wavefronts to converge to a point in the focal plane of the transmitter mirror. Both beam waists could then be co-located, by adjustment, with the target image. The beams diverge rapidly after the focal point, reaching 1/e diameters of approximately 17 cm by the time they reach the transmitter mirror (see Figure 5.6(c)). A second focal point is then predicted by the beam propagation equations and this occurs at an approximate distance of 900 m from the transmitter mirror (see Figure 5.6(d)). By the time the beam reaches the target at 1800 m, the beam diameter has returned to its original value of 17 cm.

2.2.1 The Optical Transceiver

An existing co-axial transceiver arrangement was employed which is illustrated in the schematic of Figure 2.10 [after Green, ref. 10]. Details of the original design constraints are available in reference [10] but essentially it was intended to serve as the transceiver for systems employing both direct and heterodyne detection. The transmitter telescope is intended to produce a spatially confined beam with a dimension at any range equal to the far field image of the detector active surface area, as imaged through the receiving telescope. A co-axial system has the advantages of simplifying system alignment and of isolating the detector from the transmitted pulse until after it has left the transceiver. No provision was made for steering the entire telescope assembly but adjusting screws were included in the mounting of each mirror for minor corrections to the imaged field of view.

The transmitter and receiver mirrors have diameters of 15 cm and 30 cm respectively, with respective focal lengths of 120 cm (f/8) and 180 cm (f/6). Both are arranged in a Newtonian configuration. Located behind the transmitter mirror is a large, 15 cm diameter Newtonian flat, which is used to direct the focused radiation, collected by the receiving mirror, out of the telescope casing and onto subsequent optics below. In front of the transmitter telescope is another Newton flat, 2.5 cm in diameter, used to reflect the diverging laser beams onto the transmitter mirror. Figure 2.11 is a view of the co-axial transceiver from outside the building showing both mirrors, the spider mount for the 2.5 cm flat, and the internal baffle surrounding the transmitter beam path to provide optical isolation between the two telescopes. Both the internal and external baffles were extended to a shuttered opening in the door of the laboratory. Discernible below the telescopes, and their supporting cross beams (attached to the roof), is the primary optical rail with the vertical section carrying the Tx lens visible in the foreground.

2.2.2 Alignment of the Transmitter Telescope

The lidar system target is illustrated in the photograph of Figure 5.5. This is an advertising sign, painted on a corrugated surface which forms part of the superstructure of a building belonging to a paint manufacturer. It is 1.8 km distant from the lidar site, approximately 30 m from the ground, and is angled away from the beam axis. Figure 2.12 is a reproduction of a section of an O.S. map (1:25000, first series, sheet TA 03) showing the lidar beam paths between University complex and the paint works. It will be noticed that gasometers and other industrial structures lie in close proximity to the beam path.

The "o" in the word "Croda" proved to be a very convenient target for alignment purposes. The exact dimensions of this letter are unknown but it is estimated to be approximately 1 m in diameter. Using the thin lens formula for transverse magnification,

$$M_T = \frac{S'}{S} = \frac{f}{s-f} \tag{2.38}$$

where S' is the distance from the lens, or mirror, to the object, S is the distance to the image, and f is the focal length, the size of the target in the image plane of the transmitter telescope is found to be $\sim$ 0.7 mm (f = 1.2 m, S = 1800 m). By placing a paper disc in the image plane, the target image could be located. Initially, the output from a HeNe laser was propagated through to the 90% beam splitter and Tx lens, until the beam axis was central to all components and the focused HeNe beam coincided with the target image. This procedure was then repeated with each low pressure CO_2 gain section operating in turn, but with the added complexity of the 50% beam splitter installed. The location of the Tx lens was adjusted along the beam axis until a sharp "burn through" hole appeared within the imaged letter "o" of the target.

In terms of geometric optics, the target and its image would form a set of conjugate focal points, but propagating Gaussian beams obey different reciprocity relationships. These, however, affect only the relative locations of the beam waists and focal points, and not the position where the optic axis intercepts the target. The above alignment procedure is therefore adequate for ensuring the correct locations of the optical axis out to the target. Differing spot sizes are to be expected at the target, in any event, due to the differing propagation paths between laser 1, laser 2 and the Tx lens.

2.2.3 Receiver Alignments

The essential requirement here is to focus the energy collected by the receiver mirror onto the active area of a nitrogen cooled Lead Tin Telluride detector (LTT). This is illustrated schematically in Figure 2.10, and the siting of the components is depicted in Figures 2.7(b) and 2.7(c). Using (2.38), the transverse dimension of the ∿1 m target in the image plane of the receiver is 1 mm since M_t (Receiver Telescope) = 10^{-3} (f = 1.8 m and s = 1800 m). The size of the Airy disc in the image plane, due to the entrance pupil formed by the receiver mirror, is given by

$$d = 2.44 \frac{R\lambda}{D} \qquad (2.39)$$

where R is the distance from the centre of the entrance pupil to the edge of the disc and D is the entrance pupil diameter. For the receiver mirror, therefore, d ≅ 150 μm which is comparable to the diameter of the active area of the detector (110 μm).

However, to simplify the alignment of the target image with the detector element, a zinc selenide lens (f = 25 mm) was installed between the focal plane of the receiver telescope and the detector. The converging rays from the receiver mirror were allowed to pass through the focal point and diverge again before encountering the lens at a distance, S, of 50 mm from the focal point. Application of the simple lens formula

$$\frac{1}{f} = \frac{1}{S} + \frac{1}{s'} \qquad (2.40)$$

gives the new image distance, S', from the "Rx lens", also as 50 mm (see Figure 2.13). This configuration gives an extra transverse magnification of unity (using 2.38), so the image size remains the same.

It will be noticed from Figure 2.7(b) and 2.7(c) that the reflecting surface used to turn the converging rays from the receiver Newtonian flat along the primary optical rail, is not a mirror but a 90% (germanium) beam splitter. By placing an iris at the focal point of the receiver mirror (see Figure 2.13), and allowing the CW output from one of the low pressure gain sections to propagate along the primary optical rail via both 90% beam splitters (see Figures 2.7(a) and 2.7(b)), a point source of 10 μm radiation is made available at the iris aperture, having a larger irradiance than the 10 μm background. During the alignment procedure a chopper wheel was installed in the path of the CW beam to provide a low frequency ac signal at the detector. Detector alignment, along all three axes, was accomplished by maximising this signal.

RÈFERENCES

[1] MENYUK, N., KILLINGER, D.K. and MENYUK, C.R. "Limitations of signal averaging due to temporal correlation in laser remote-sensing measurements", Applied Optics, 21(18), Sept. 1982, pp. 3377-3383.

[2] JUDD, O.P. and WADE, J.Y. "Investigations of UV Preionized Electrical Discharge and CO_2 Laser", IEEE Journal of Quantum Electronics, QE10(1), Jan. 1974.

[3] ANDREWS, K.J., DYER, P.E. and JAMES, D.J. "A rate equation model for the design of TEA CO_2 Oscillators", J. of Physics E., 8, 1975, pp. 493-497.

[4] BLOOM, A.L. "Gas Lasers", John Wiley and Sons, 1968.

[5] LENGYEL, B.A. "Lasers" (2nd Edition), John Wiley and Sons, 1968.

[6] SIEGMAN, A.E. "An Introduction to Lasers and Masers", McGraw-Hill, 1971.

[7] YARIV, A. "An Introduction to Optical Electronics", Holt, Rinehart, and Winston, Inc., 1971.

[8] BOYD, G.D. and GORDON, J.P. "Confocal multimode resonator for millimeter through optical wavelength masers", Bell Syst. Tech. J., 40, pp. 489-508, March 1961.

[9] BOYD, G.D. and KOGELNIK, H. "Generalised confocal resonator theory", Bell Syst. Tech. J., 41, July 1962, pp. 1347-1369.

[10] GREEN, B.A. "Coherent Laser Radar for Atmospheric Sensing", Ph.D. Thesis, Dept. of Applied Physics, Hull University, England, 1979.

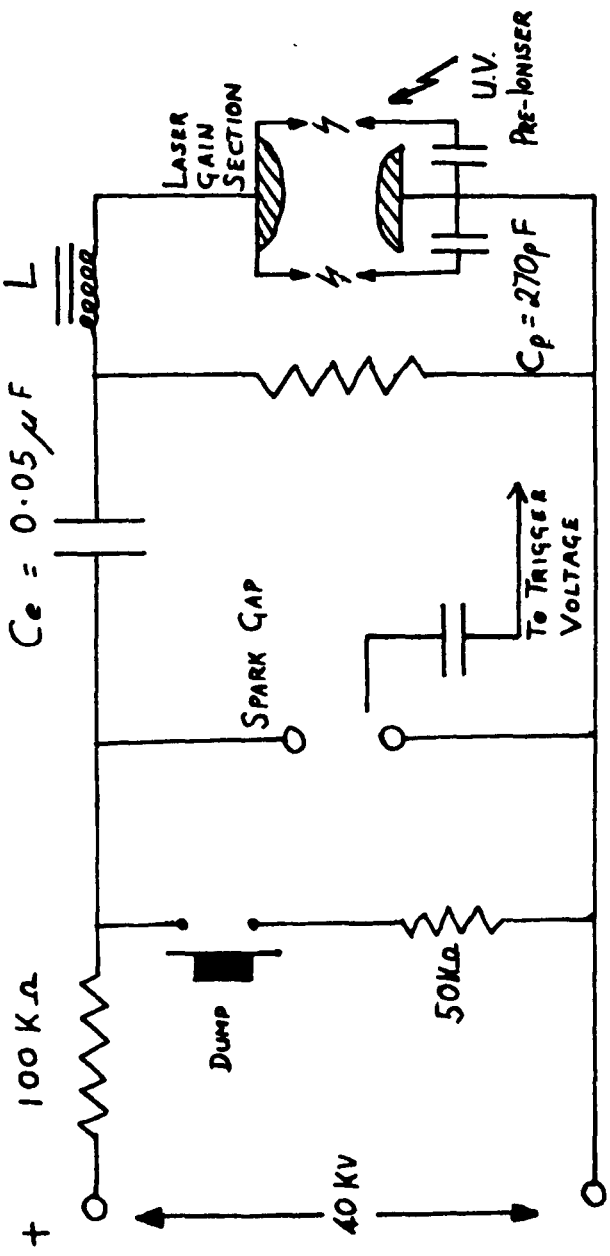

LASER GAIN SECTION
U.V. PRE-IONISER
L
1000R
C_p = 270pF
C_e = 0·05μF
SPARK GAP
TO TRIGGER VOLTAGE
100 KΩ
DUMP
50KΩ
+
40 KV

Fig 2.1 (a)

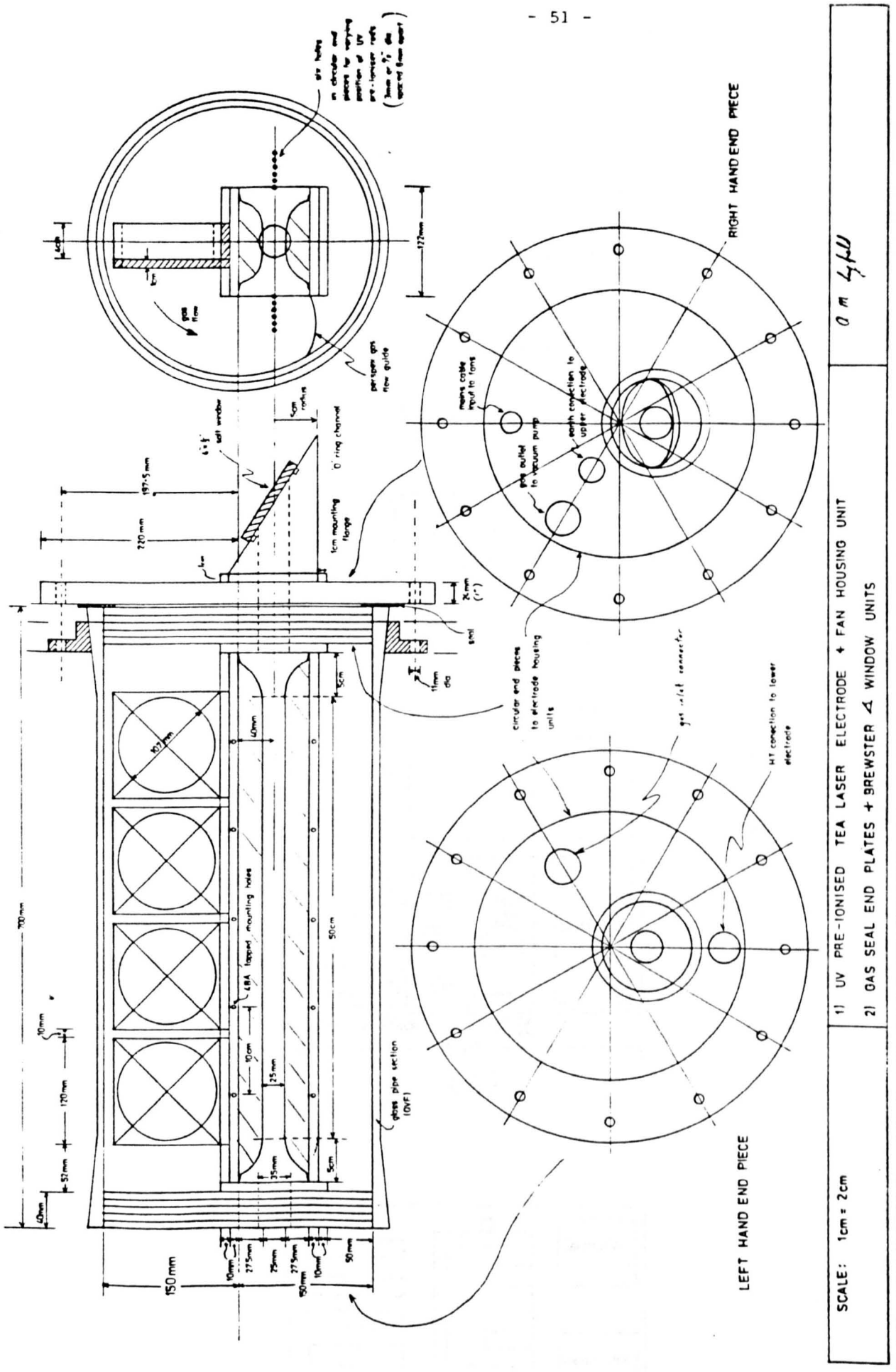
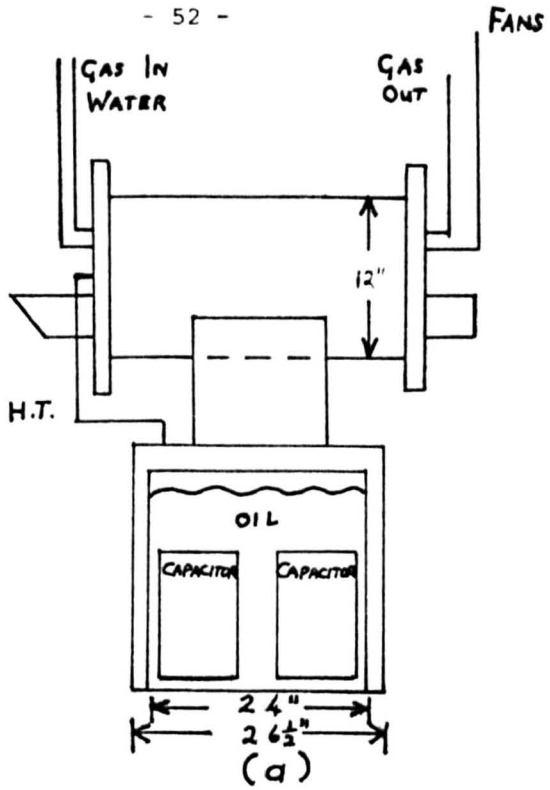
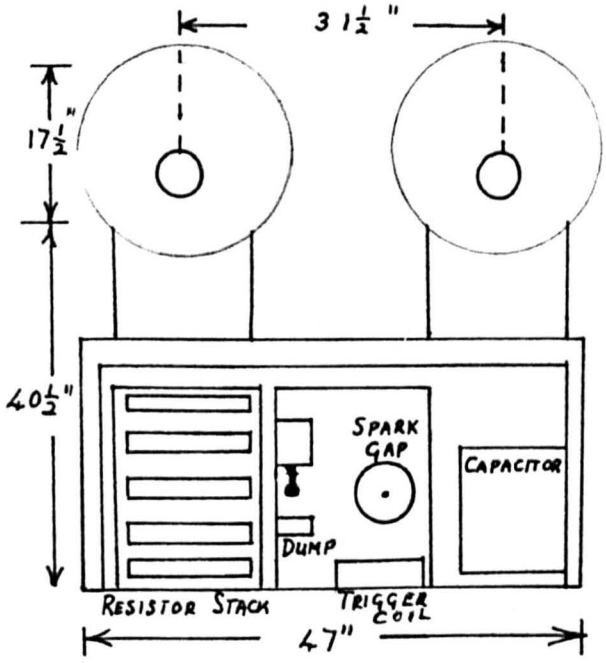

RIGHT HAND END PIECE
LEFT HAND END PIECE
SCALE: 1cm = 2cm
1) UV PRE-IONISED TEA LASER ELECTRODE + FAN HOUSING UNIT
2) GAS SEAL END PLATES + BREWSTER ∠ WINDOW UNITS
0 m Lyttl
gas flow
gas outlet to vacuum pump
mains cable input to fans
earth connection to upper electrode
circular end pieces to electrode housing units
gas inlet connector
HT connection to lower electrode
1cm mounting flange
O ring channel
salt window
glass pipe section (OVF)
4BA tapped mounting holes
50cm
700mm
150mm

Fig. 2.1 (b)

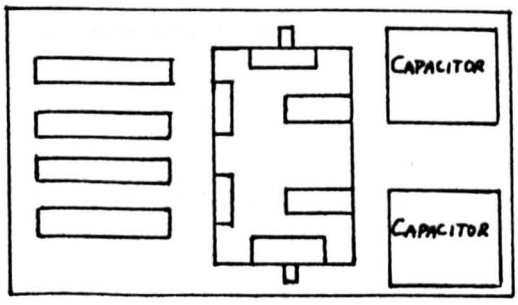

FANS
GAS IN WATER
GAS OUT
12"
H.T.
OIL
CAPACITOR
CAPACITOR
24"
26½"

(a)

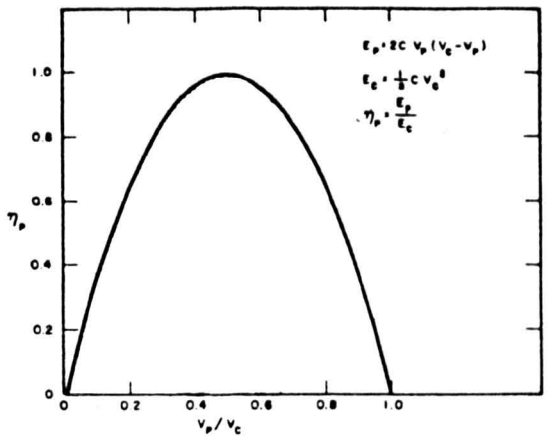

31½"
17½"
40½"
SPARK GAP
CAPACITOR
DUMP
RESISTOR STACK
TRIGGER COIL
47"

(b)

Fig. 2.2

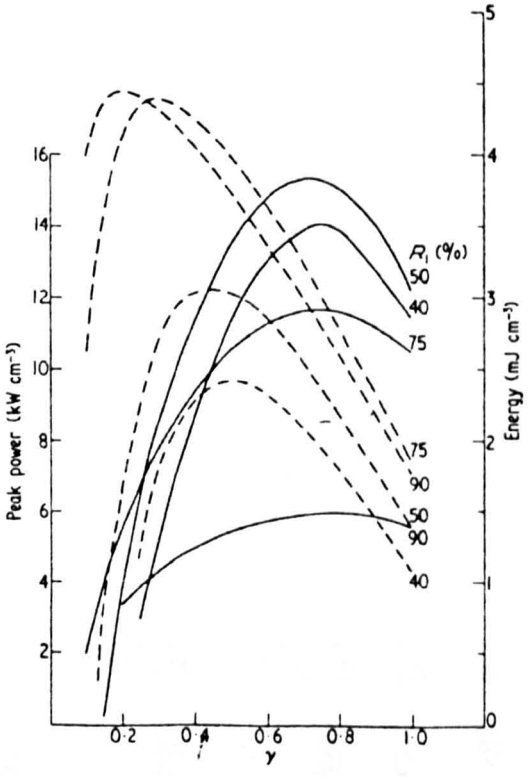

CAPACITOR
CAPACITOR

(c)



Single-section pulse-forming network transfer efficiency as a function of V_p/V_c. Theoretically, a complete transfer of energy into the plasma is obtained only for the condition that $V_p/V_c = 0.5$.

Fig. 2.3 (after Judd and Wada, 1974)



Laser energy (broken line) and peak power (full line) as a function of molecular gas ratio γ for various output coupling mirror reflectivities R_1. $R_2 = 97{\cdot}5\%$, $A = 1\%$.

Fig. 2.5 (after Andrews, Dyer and James, 1974)

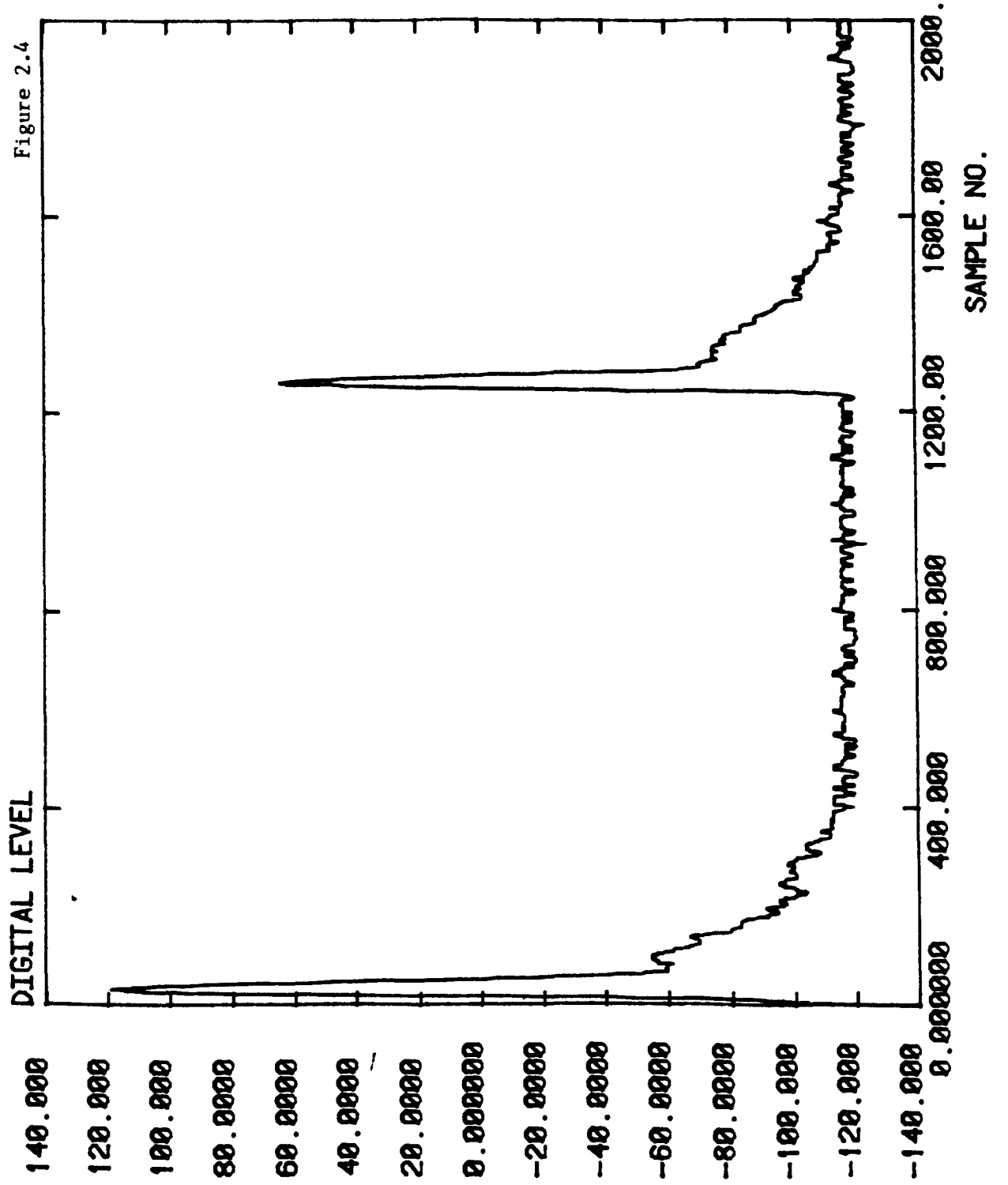

SIFT3.007
Transmitted and Received Pulses
DIGITAL LEVEL
140.000
120.000
100.000
80.0000
60.0000
40.0000
20.0000
0.000000
-20.0000
-40.0000
-60.0000
-80.0000
-100.000
-120.000
-140.000
0.000000
400.000
800.000
1200.00
1600.00
2000.00
SAMPLE NO.

Figure 2.4

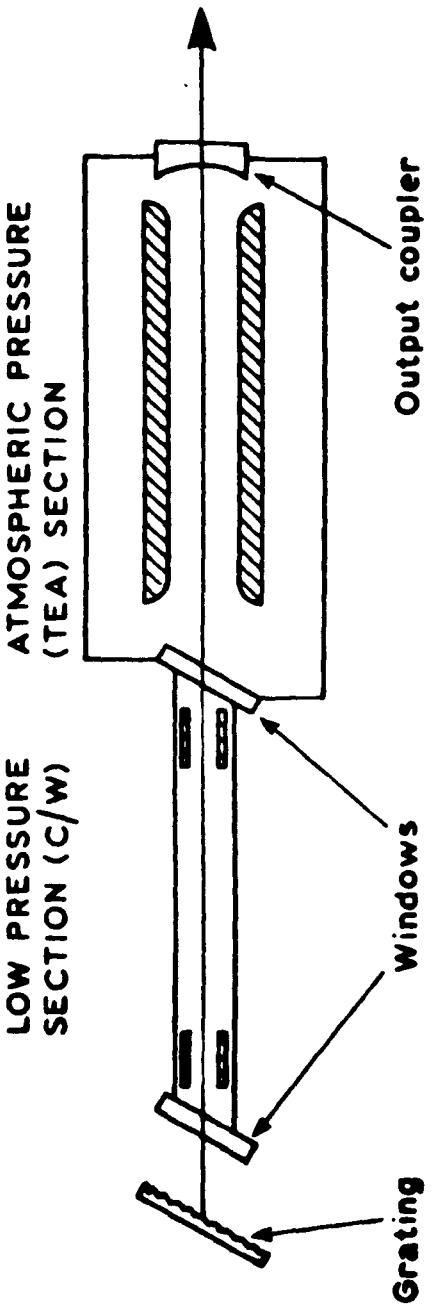

ATMOSPHERIC PRESSURE
(TEA) SECTION
LOW PRESSURE
SECTION (C/W)
Output coupler
Windows
Grating

Fig. 2.6

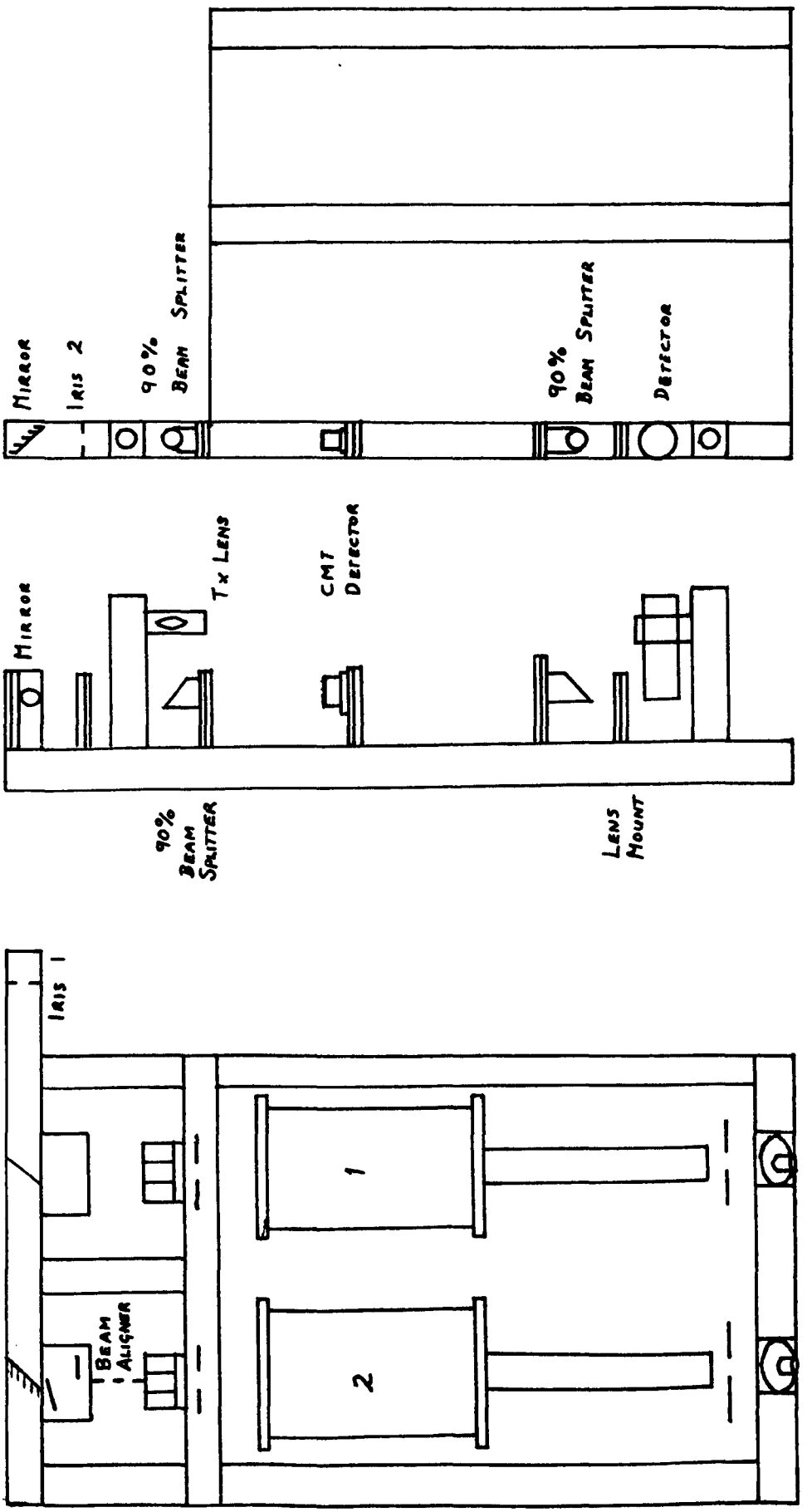

Mirror
Iris 2
90% Beam Splitter
90% Beam Splitter
Detector
Mirror
Tx Lens
CMT Detector
90% Beam Splitter
Lens Mount
Iris 1
Beam Aligner
1
2

Fig. 2.7

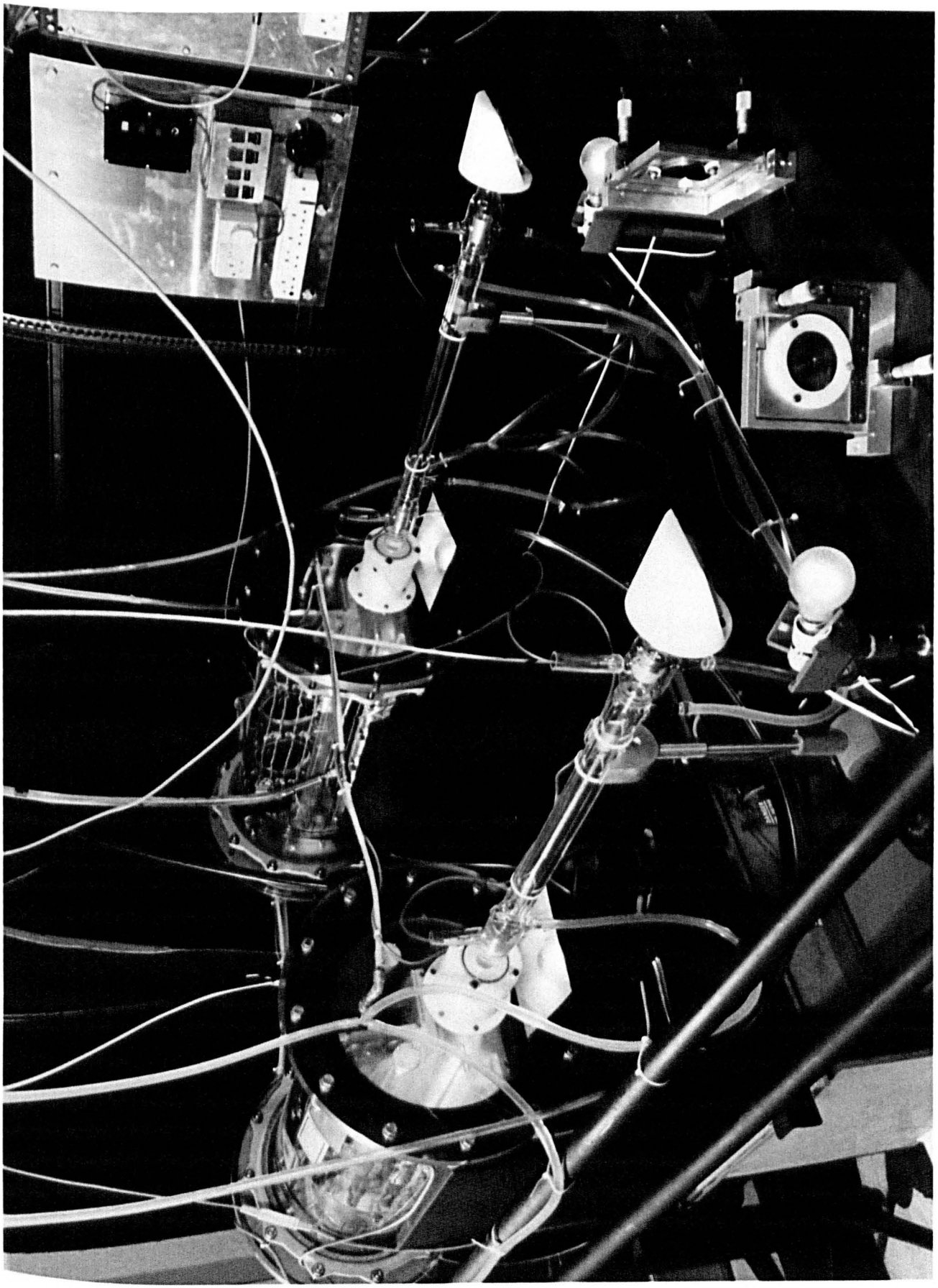
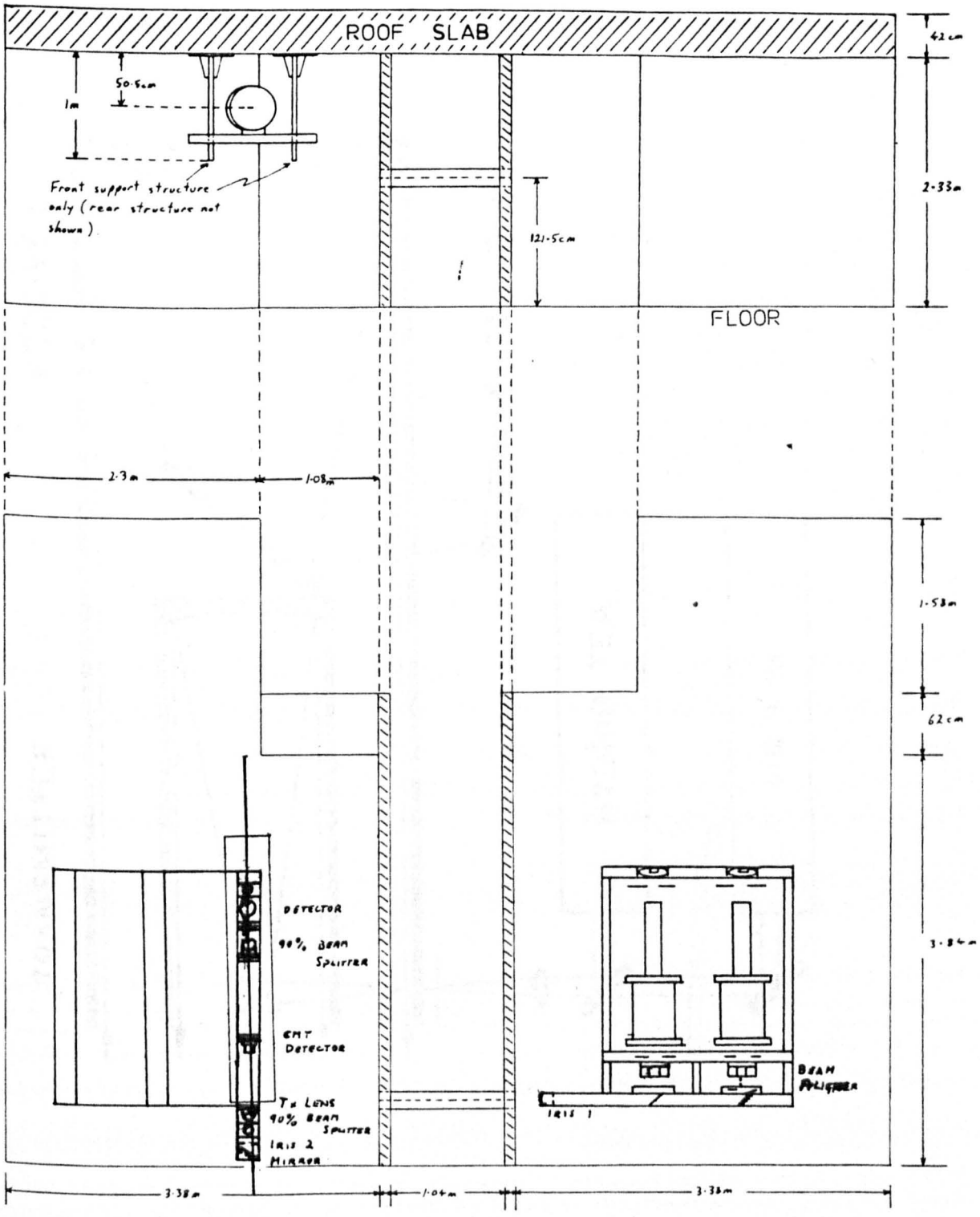


Fig. 2.8



A LAYFIELD	Location of Telescope Support Structure And Tube Through Walls (Roof Laboratory - Appl. Physics)	Scale : 1cm = 20cm

Fig. 2.9

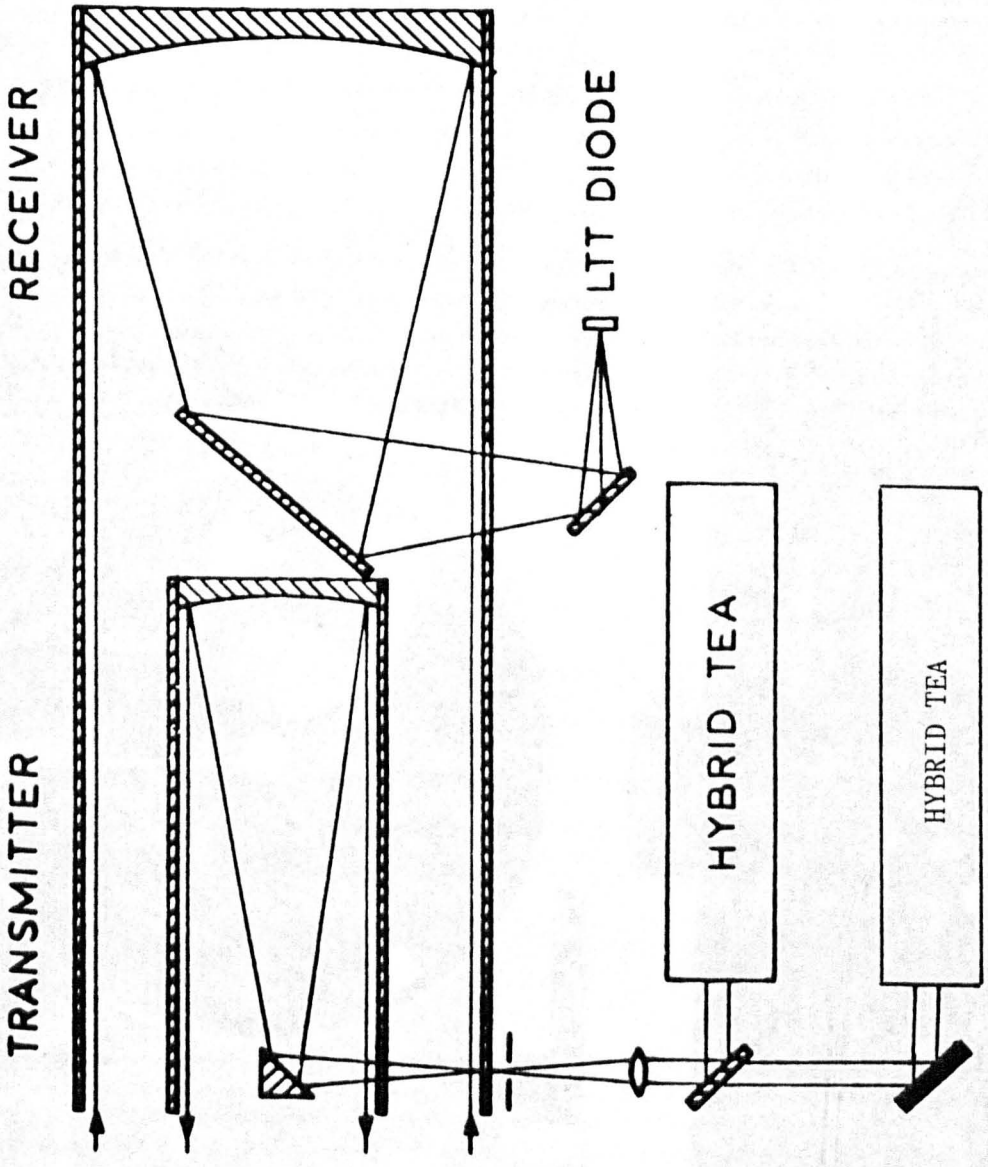


Fig. 2.10

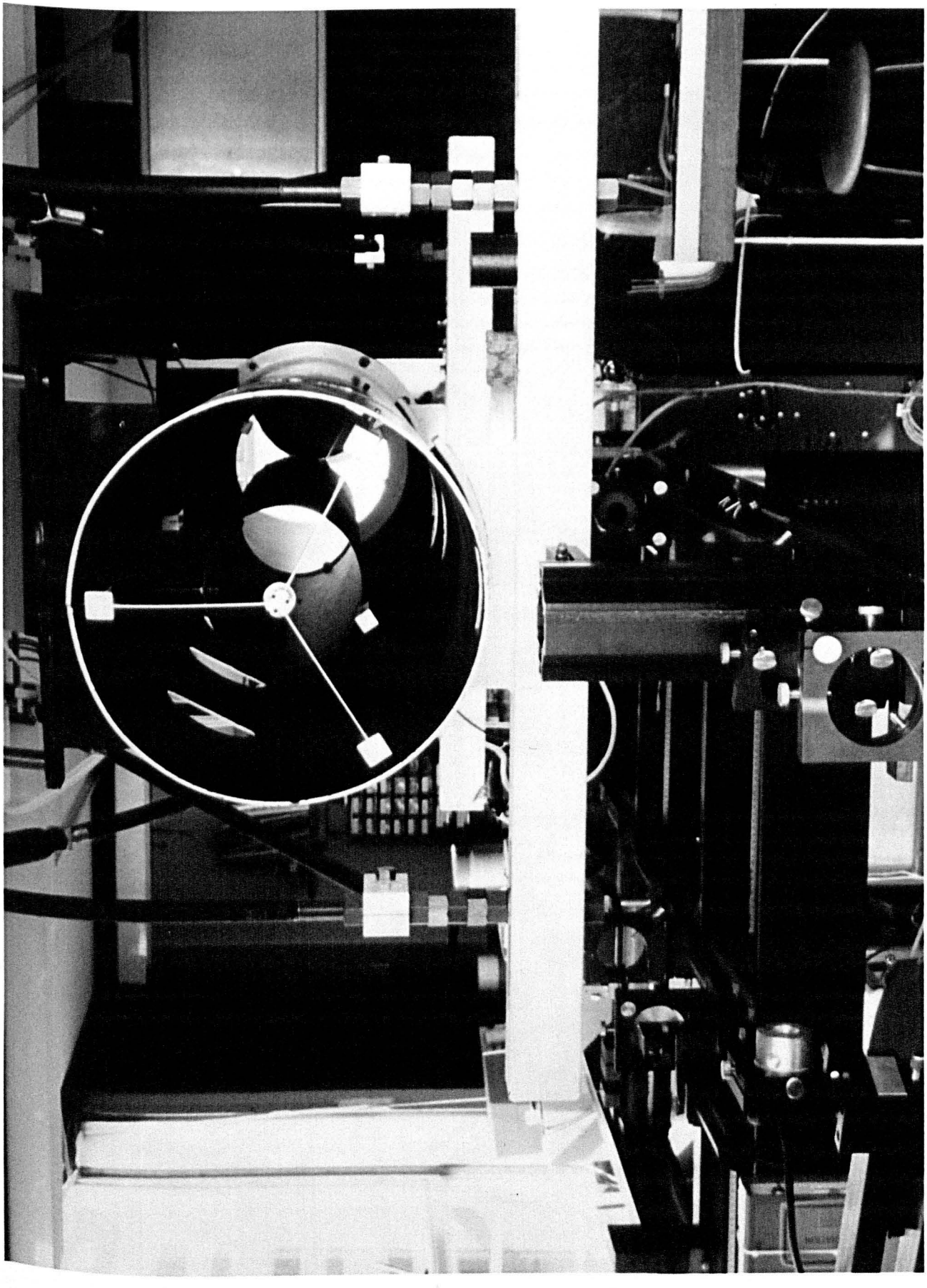


Fig. 2.11

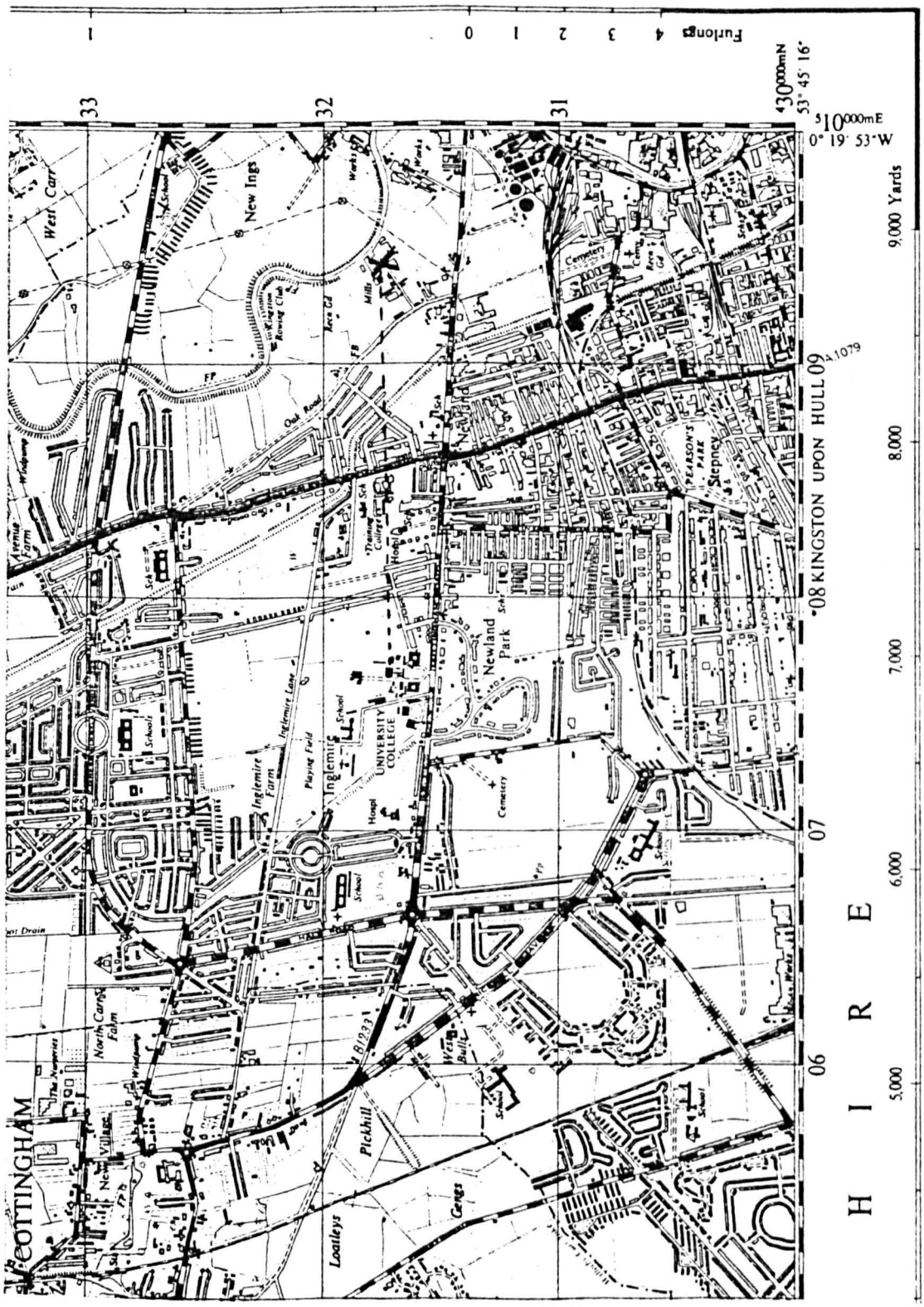

SHEET TA03
Furlongs 4 3 2 1 0 1
5 10 000mE
0° 19' 53"W
430 000mN
53° 45' 16"
33
32
31
08 KINGSTON UPON HULL 09
07
06
H I R E
5,000
6,000
7,000
8,000
9,000 Yards
NOTTINGHAM
New Ings
West Carr
Inglemire Farm
Inglemire Lane
Playing Field
Inglemire
UNIVERSITY COLLEGE
Hospl
Newland Park
Cemetery
Pickhill
Loaleys
Gengs
North Carrs Farm
Oak Road
Stepney
PEARSON'S PARK
Made and published by the Director General of the Ordnance Survey, Southampton.
Reprinted with the addition of new major roads.

Fig. 2.12

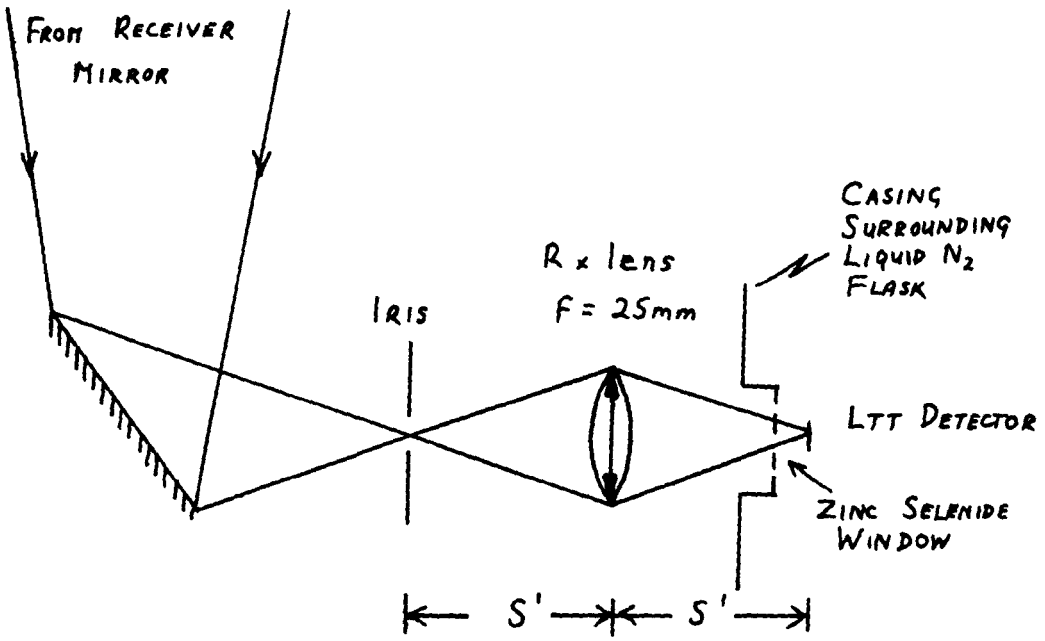

From Receiver
Mirror
Iris
R x lens
f = 25mm
Casing
Surrounding
Liquid N_2
Flask
LTT Detector
Zinc Selenide
Window
S'
S'

Fig. 2.13

CHAPTER 3.0

LIDAR SYSTEM PARAMETERS

3.1 The Lidar Equation

The Beer-Lambert-Bouger Law of expression (1.3) is reproduced here using the new notation,

$$\frac{d\phi(r)}{dr} = -\alpha\ \phi(r) \tag{3.1}$$

and in this form relates the decay of the transmitted photon flux, $d\phi(r)$, over a propagation path length, dr, to the instantaneous flux, value $\phi(r)$, and a constant, α, characteristic of the attenuating medium. α is, in fact, the total atmospheric extinction coefficient and since the flux, $\phi(r)$, is equivalent to power $/h\nu$, where h is Planck constant and ν the optical frequency, (3.1) may be re-cast into the form

$$\frac{dP(r)}{dr} = -\alpha\ P(r) \tag{3.2}$$

in which power replaces flux.

Integration of (3.2) yields

$$P(r) = P_T e^{-\alpha r} \tag{3.3}$$

where P(r) represents the power remaining at range r and P_T, the original transmitted power. To obtain the power backscattered from a remote topographic target at range, r, certain assumptions must be made regarding the reflectivity of a non-cooperative target. The simplest assumption has the radiation scattered isotropically throughout 2π steradians so that if the target reflectivity is, ρ', the power per unit solid angle scattered back to the transmitter site will be

$$P'(2r) = P_T \frac{\rho'}{2\pi r^2} e^{-2\alpha r} \tag{3.4}$$

It is nearly always assumed, however, that topographic targets in general are not isotropic scatterers but possess Lambertian scattering properties in which the scattering parameter becomes ρ/π instead of $\rho'/2\pi$ so that (3.4) must be modified to [1,2],

$$P'(2r) = P_T \frac{\rho}{\pi} \frac{e^{-2\alpha r}}{r^2} \tag{3.5}$$

If a receiving telescope of mirror area, A_R, is also located at the site of the transmitter then the maximum power it will be capable of collecting is

$$P_R = P_T \frac{\rho}{\pi} \frac{A_R}{r^2} e^{-2\alpha r} \tag{3.6}$$

This expression can be further generalised to include a factor C, with a value between 0 and 1, which represents a measure of the optical efficiency of the transceiver. Furthermore, the atmospheric extinction coefficient, α, is usually regarded as being variable between the lidar site and target so that the argument to the exponential terms must be modified also. The final form of the lidar equation for Lambertian scattering from a topographic target is therefore,

$$P_R(r) = P_T C \frac{\rho}{\pi} \frac{A_R}{r^2} \exp\left[-2 \int_0^r \alpha(r)dr\right] \tag{3.7}$$

It is worth back tracking at this point to consider the power back scattered from the atmosphere itself. If the number of scattering particles per unit volume is n_S, and each particle has a scattering cross section of σ_S, then the total scattering cross section per unit volume of the atmosphere is $n_S\sigma_S$, which has the dimension $[L^{-1}]$. In order to obtain the amount of radiation scattered out of the beam $n_S\sigma_S$ must be multiplied by ρ_S which has the dimension [L] and is related to

the spatial extent of the laser output along the propagation path. The product $n_S\sigma_S\rho_S$ is clearly dimensionless (and increases the attenuation of the beam energy) so that, for the energy scattered from the beam path at range, r, (3.3) becomes,

$$P_S(r) = P_T n_S \sigma_S \rho_S \, e^{-\alpha r} \tag{3.8}$$

A physical interpretation for the quantity, ρ_S, may be obtained by first noting that, for a pulsed laser system, only a relatively small volume of the atmosphere along the beam axis will contribute to the scattered radiation at any instant in time. If the maximum range from which scattered radiation can be received is

$$r = \frac{ct}{2} \tag{3.9}$$

where t = 0 at the time of transmission of the pulse, then assuming the pulse itself is of duration t_p, the minimum range from which radiation is received at time t is

$$r' = \frac{c}{2}(t - t_p) \tag{3.10}$$

Scattering centres within the range interval,

$$r - r' = \frac{ct_p}{2} = \rho_S \tag{3.11}$$

will therefore all contribute to the total radiation removed from the laser pulse (see Figure 3.1).

ρ_S is therefore the "effective pulse length", or depth of the scattering volume, and is only half the actual pulse length. Expression (3.8) therefore becomes

$$P_S(r) = P_r n_S \sigma_S \, \frac{ct_p}{2} \, e^{-\alpha r} \tag{3.12}$$

Assuming this power to have been isotropically scattered into 4π steradians, the power returning back over the range r, and collected by a receiver mirror of area A_R, will be

$$P_R(2r) = P_T n_S \sigma_S \frac{ct_p}{2} \frac{A_R}{4\pi r^2} e^{-2\alpha r} \qquad (3.13)$$

where the argument of the exponential term has been doubled to account for the outgoing and return path length of $2r$ between the scattering volume and the receiver.

A further parameter, C, must be included in the above expression, having a value in the range $0 < C < 1$, to allow for the fractional losses associated with real transceiver optics. Also, if the exponential term is generalised in recognition of the range and wavelength dependence of absorption then, with some re-grouping of terms, (3.13) becomes

$$P_R = P_T C \left[\frac{ct_p}{2}\right]\left[\frac{n_S \sigma_S}{4\pi}\right] \frac{A_R}{r^2} \exp\left[-2\int_o^r \alpha(r, \lambda)\, dr\right]$$

$$= P_T C \left[\frac{ct_p}{2}\right] \beta \frac{A_R}{r^2} \exp\left[-2\int_o^r \alpha(r,\lambda)\, dr\right] \qquad (3.14)$$

where $\beta = (n_S \sigma_S / 4\pi)$ is the atmospheric backscatter coefficient. β is defined as the fractional amount of incident energy scattered, per unit solid angle, in the backward direction, per unit atmospheric length, and may vary both as a function of time and the target volume range. It is also dependent on wavelength and scattering particle size distributions.

The lidar equation for atmopheric scattering is valid as long as the BLB law can be assumed to hold true and, in practice, this translates into an assumption that all of the scattered energy is permanently removed from the transmitted pulse. In reasonably transparent atmospheres, and for CO_2 laser wavelengths, this is usually the case [2] but in cloud, fog or thick haze, optical scattering is much more enhanced and some of this multiply scattered radiation can actually re-enter the beam, necessitating a modification of (3.14).

Both lidar equations, (3.7) and (3.14) can now be used to give an approximate indication of the relative powers backscattered to the receiver from the two different types of target. By forming the ratio of the received powers;

$$\frac{P_R \text{ (topographic target)}}{P_R \text{ (atmospheric target)}} = \frac{2\rho}{ct_p\beta\pi} \qquad (3.15)$$

and assuming $\rho = 0.1$ (a conservative estimate), $\beta = 10^{-7}sr^{-1}m^{-1}$ (a typical value at CO_2 wavelengths) and a pulse duration of 5×10^{-7} seconds (true for the Hull Lidar System), it is found that

$$\frac{P_R \text{ (solid target)}}{P_R \text{ (atmospheric target)}} \simeq 4 \times 10^3$$

Variation of at least an order of magnitude can be expected either side of this value depending on the precise values of β and ρ.

Clearly, however, received powers of at least two orders of magnitude greater than those due to atmospherically scattered radiation can be expected from any given range using a topographic target.

Unless the transmitted energies used by direct detection CO_2 lidar systems are at least of the order of several Joules, range resolved measurements tend to be impractical over any useful path length due to the background radiation limited performance of currently available infra-red detectors operating near 10 μm. Beyond transmitted energies of 5-10 Joules, eyesafe radiation levels become an important consideration, even though the cornea is opaque to optical radiation near 10 μm and therefore tends to limit the potential for damage compared to systems using visible wavelengths. It is possible that developments in infra-red detection

techniques, such as the frequency up-conversion method proposed by Itabe and Bufton [1], may eventually improve the range resolved measurement capabilities of direct detection,for 10μmwavelength systems restricted to transmitted energies of less than 5 Joules.

3.2 Absorption Spectra

The absorption term in the lidar equation (3.7) was first discussed in section 1.3.3 and was presented as consisting of four terms;

$$A = \alpha_{AM} + \alpha_{Ap} + \alpha_{SR} + \alpha_{SMIE} \qquad (1.2)$$

where α_{AM} and α_{Ap} are the absorption components due to molecular absorption and particle absorption respectively, and α_{SR} and α_{SMIE} are the scattering components due to the Rayleigh and Mie processes. All α terms are of dimension L^{-1}. Although molecular absorption is of primary interest in this section, the remaining three terms on the right hand side of (1.2) are obviously important. Various pairs of CO_2 emission lines have been recommended [3] to minimise the differential absorption components contributed by those terms and also by interfering molecular species which have absorption spectra overlapping absorption features of the molecular species of interest. Advantages of the differential absorption technique will be discussed in the next section but will be restricted to the case where two probe wavelengths are used to form the power ratio estimator. This conforms to the experimental restrictions relevant to the data obtained and analysed later in this work. It is anticipated that future developments in lidar, which strive towards higher pollutant or trace gas concentration accuracies, may require the near simultaneous transmission of more than two wavelengths in order to resolve absorption components due only to the molecular species of interest.

Since many gas molecules have a vibrational-rotational absorption band between 2 and 15 μm [4], CO_2 gas lasers with emission lines between 9 and 11 μm appear to be particularly suited to the task of probing the concentrations of atmospheric pollutants. The pulsed transversely excited gain sections of the CO_2 lasers used as the source of transmitted radiation in the Hull lidar system were operated at atmosphetic pressure and therefore have emission spectra which are often comparable in line widths with the line widths of foreign or naturally occurring atmospheric constituents.

At low pressures, typically below 100 mbar, Doppler shift is the dominant contribution to the line shape and results in a Gaussian distribution for the absorption cross section per molecule of the form [5]

$$K_m(\nu) = \frac{\alpha_m(\nu)}{n_m} = \frac{2\,S(\ln 2)^{\frac{1}{2}}}{\Delta\nu_D} \exp\left[\frac{-4(\nu - \nu_o)^2 \ln 2}{(\Delta\nu_D)^2}\right] \qquad (3.16)$$

where $K_m(\nu)$ is typically given in units of cm^2, ν is the optical frequency (cm^{-1}), ν_o is the line centre frequency (cm^{-1}), n_m is the number of molecules of a particular species per unit volume (cm^{-3}),

$$\Delta\nu_D \simeq \left(\frac{214}{\lambda}\right)\left(\frac{T}{M}\right)^{\frac{1}{2}} \text{ MHz} \qquad (3.17)$$

is the full width at half maximum (cm^{-1}), M is the molecular weight [6] and

$$S(T) = \frac{S(T_o)\,Q_v(T_o)\,Q_r(T_o)}{Q_v(T)\,Q_r(T)} \exp\left[\frac{1.439\,E''\,(T - T_o)}{T\,T_o}\right] \qquad (3.18)$$

is the integrated line intensity (cm) [7]. Q_v and Q_r are the vibrational and rotational partition functions and T_o is usually taken to be 296 K.

Such a line shape is convolved with what would be, in the absence of Doppler shift , the natural line width of a molecule due to the finite life time of the radiative transition. Beyond 100 mbar, the line width increases monotonically due to collisions between molecules of the same species (self broadening) or some other species, and eventually collisional broadening (or "pressure broadening") dominates over doppler broadening to produce a line shape referred to as a "Lorentzian". The governing relationship for the absorption cross section in this case is [6],

$$K_m(\nu) = \frac{\alpha_m(\nu)}{n_m} = \frac{S\ \Delta\nu_p}{\pi(\nu - \nu_o)^2 + \Delta\nu_p^2} \tag{3.19}$$

where

$$\Delta\nu_p = \Delta\nu_{p_o} \left(\frac{T_o}{T}\right)^{\frac{1}{2}} \left(\frac{p}{p_o}\right) \tag{3.20}$$

is the line width at temperature, T, and pressure, P. T_o is taken to be 296 K and P_o is usually equated to 1013 mbar [7].

Clearly the magnitude of the absorption cross section at any given wavelength is dependent on both temperature and pressure and must therefore be regarded as a potential source of error if, as is likely, the exact values of the temperature and pressure are unknown over the measurement path.

When a large number of absorption lines, belonging to many different molecular species, contribute to the total absorption at any specific wavelength, the total absorption coefficient becomes

$$\alpha_{m_{TOTAL}}(\nu) = \sum_j \sum_i \frac{S_{ij}\ \Delta p_{ij}}{\pi((\nu - \nu_o)^2 + \nu_{p_{ij}}^2)} \cdot n_j \tag{3.21}$$

where n_j is the molecular abundance of the jth molecular species, and i denotes each line belonging to the jth species.

3.3 The Differential Absorption Technique

The path integrated lidar equation, (3.7), of Section 3.1 can be re-written as

$$P_R = P_T \frac{\rho}{\pi} \frac{A_R}{r^2} C \exp[-2A(\lambda)] \tag{3.22a}$$

where

$$A(\lambda) = \int_o^r \alpha(r,\lambda)dr \tag{3.22b}$$

is referred to as the (one way) absorbance.

There are two such expressions, however, for a dual wavelength lidar system;

$$P_R(\lambda_1) = P_T(\lambda_2) \frac{\rho_1}{\pi} \frac{A_R}{r^2} C_1 \exp[-2A(\lambda_1)] \tag{3.23a}$$

$$P_R(\lambda_2) = P_T(\lambda_2) \frac{\rho_2}{\pi} \frac{A_R}{r^2} C_2 \exp[-2A(\lambda_2)] \tag{3.23b}$$

For the Hull lidar system $P_T(\lambda_1)$ and $P_T(\lambda_2)$ are transmitted with a small temporal separation, of the order of 50 microseconds, permitting the two received powers $P_R(\lambda_1)$ and $P_R(\lambda_2)$ to be resolved at a single detector (see Chapter 4).

The two target scattering coefficients, ρ_1 and ρ_2, may be expected to vary on a shot to shot basis due to instabilities in the system alignment and scintillation effects which will cause different areas of the target to be illuminated. However, the optical transmission paths for the two wavelengths were carefully aligned prior to each measurement sequence (10,000 shots) and the alignment checked by inserting a paper disc into the focal plane of the transmitting telescope. A series of calibration pulses punctured the paper disc, thus making it possible to check that not only did both beam paths

come to a focus at the same point, but also that they were coincident with the centre of the target image. The alignment could therefore also be checked throughout the measurement sequence and, for the purposes of this section, it is consequently assumed that, on average, $\rho_1 = \rho_2 = \rho$.

If λ_1 corresponds to the "reference wavelength" (ideally, not absorbed by any atmospheric constituent) and λ_2 corresponds to the "measurement of wavelength" (absorbed only by the molecular species of interest) then, forming the ratio of 3.23(a) and (b) gives,

$$\frac{P_R(\lambda_1)}{P_R(\lambda_2)} = \frac{P_T(\lambda_1)}{P_T(\lambda_2)} \frac{C_1}{C_2} \exp\left\{-2[A(\lambda_1) - A(\lambda_2)]\right\}$$

or

$$\frac{P_R(\lambda_1)\, P_T(\lambda_2)\, C_2}{P_R(\lambda_2)\, P_T(\lambda_1)\, C_1} = \exp\left\{-2[A(\lambda_1) - A(\lambda_2)]\right\}$$

which yields

$$\Delta A = A(\lambda_2) - A(\lambda_1) = \frac{1}{2} \log_e \left[\frac{P_R(\lambda_1)\, P_T(\lambda_2)\, C_2}{P_R(\lambda_2)\, P_T(\lambda_1)\, C_1} \right] \qquad (3.24)$$

Now the absorption coefficient α_m for a particular molecular species, m, is defined as

$$\alpha_m = n_m K_m \qquad (3.25)$$

where n_m is the number of molecules of type m per unit volume (dimension, L^{-3}) and K_m is the absorption cross section per molecule (dimension, L^2). Using expression (3.23), the absorption term can be re-defined as a path integrated quantity,

$$A_m = \int_o^r \alpha_m dr = K_m \int_o^r n_m(r)dr$$

$$= K_m N_m \qquad (3.26)$$

where N_m (cm^{-2}) is the number of molecules per unit area over a path length of r. Also, the path integrated differential absorption coefficient becomes

$$\Delta A_m = A_m(\lambda_2) - A_m(\lambda_1) = N_m(K_m(\lambda_2) - K_m(\lambda_1))$$

$$= N_m K \qquad (3.27)$$

for the molecular species, m, where $K = K_m(\lambda_2) - K_m(\lambda_1)$. The concentration, γ_m, of m type molecules is defined as the number of m type molecules per unit volume (N_m) divided by the total number of atmospheric molecules per unit volume . n , so that

$$\gamma_m = \frac{n_m}{n} = \frac{N_m}{N} = \frac{\Delta A_m}{K} \cdot \frac{1}{N} = \frac{\Delta A_m}{Krn} \qquad (3.28)$$

and $N = n \int_o^r dr = nr$.

Substituting for A_m using (3.24) yields

$$\gamma_m = \frac{1}{2Krn} \log_e \left| \frac{P_R(\lambda_1)\ P_T(\lambda_2)\ C_2}{P_R(\lambda_2)\ P_T(\lambda_1)\ C_1} \right|$$

$$= \frac{1}{2Krn} \left\{ \log_e \left[\frac{P_N(\lambda_1)}{P_N(\lambda_2)} \right] + \log_e \left[\frac{C_2}{C_1} \right] \right\} \qquad (3.29)$$

where P_N denotes the normalised power, P_R/P_T. Clearly the second term on the right hand side represents an error term.

Ideally $C_1 = C_2$ and this term would disappear. In practice however it is expected that C_1 and C_2 will be different for the Hull lidar system since two laser sources are used to provide the transmitted power $P_T(\lambda_1)$

and $P_T(\lambda_2)$; different optical path lengths are therefore introduced resulting in slightly different beam waists throughout all aperturing components in the system. This type of systematic error may be overcome, however, only if the ratio C_1/C_2 can be identified or estimated.

Published details of absorption coefficients are not always given in terms of the absorption cross-section per molecule, K_m. Typically, a quantity such as α'_m will be quoted giving the absorption per atmosphere, per cm path length for the specific molecular species, m. The relationship between the absorption coefficient per cm, α_m, and α'_m is then

$$\alpha_m = \gamma_m \alpha'_m \tag{3.30}$$

Therefore, since, $K = K_m(\lambda_2) - K_m(\lambda_1)$, and using the relationship (3.25), the absorption cross-section per molecule can be re-written as,

$$K = \frac{\alpha_m(\lambda_2) - \alpha_m(\lambda_1)}{n_m}$$

$$= \gamma_m [\alpha'_m(\lambda_2) - \alpha'_m(\lambda_1)] \; / \; n_m$$

$$= \frac{\left(\frac{n_m}{n}\right) [\alpha'_m(\lambda_2) - \alpha'_m(\lambda_1)]}{n_m}$$

$$= \frac{\alpha'_m(\lambda_2) - \alpha'_m(\lambda_1)}{n}$$

$$\text{or} \quad \Delta\alpha'_m = \alpha'_m(\lambda_2) - \alpha'_m(\lambda_1) = Kn \tag{3.31}$$

Replacing the product Kn in expression (3.29) with $\Delta\alpha'_m$ gives the alternative form

$$\gamma_m = \frac{1}{2\Delta\alpha'_m r} \log_e \left[\frac{P_N(\lambda_1)}{P_N(\lambda_2)}\right] \tag{3.32}$$

where the log term containing C_1 and C_2 has been dropped.

REFERENCES

1. HINCKLEY, E.D. (Ed.). (1976). Laser Monitoring of the Atmosphere. Springer Verlag.

2. BYER, R.L. and GARBUNY, M. (1973). Applied Optics 12(7), pp. 1496-1505.

3. PETHERAM, J.C. (1981). 'A comparison between continuously tunable and line tunable CO_2 laser systems for the remote sensing of atmospheric trace gases'. Report prepared at Hull University, Department of Applied Physics, England for the National Physical Laboratory, Teddington, England.

4. HERZBERG, G. (1950). Molecular Spectra and Molecular Structure, Vols. 1 and 2. D. Van Nostrand Co., New York.

5. RAO, K.N. (1972). Molecular Spectroscopy: Modern Research. Academic Press, New York, p. 346.

6. PENNER, S.S. (1959). Quantative Molecular Spectroscopy and Gas Emissivities. Addison Wesley, Reading, Mass., U.S.A.

7. PETHERAM, J.C. (1981). Applied Optics 20(22), pp. 3941-3946.

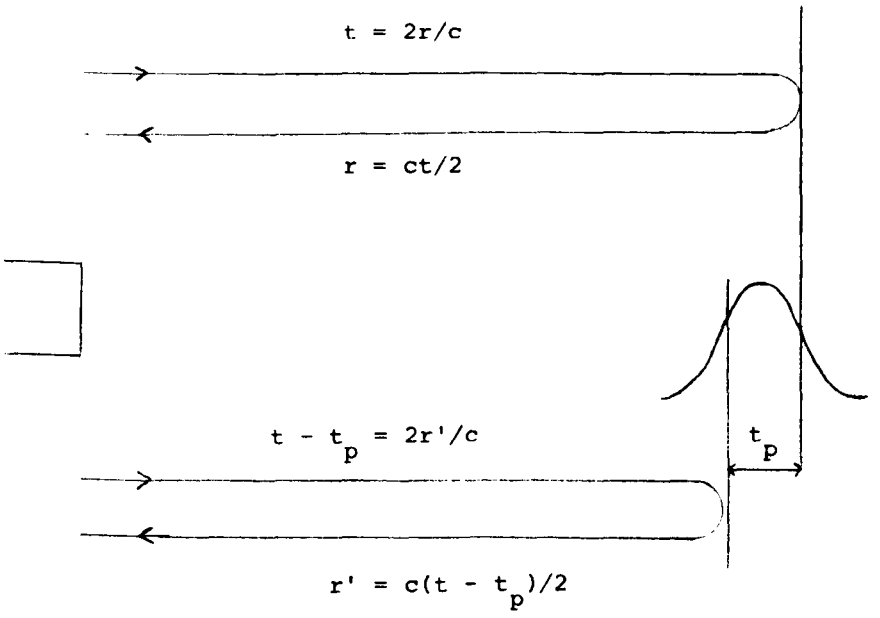


Fig. 3.1

CHAPTER 4.0

SIGNAL PROCESSING

4.1 Signal Processing Hardware

A schematic of the signal processing electronics is presented in Figure 4.1. Electrical noise problems, originating with the laser trigger and discharge pulses, significantly influenced the design of the signal processing modules. This is particularly true of the optic fibre communication channel between the telescope room and an adjacent room housing the Biomation 8100 waveform recorder and the PDP 11/10 computer. All components located in the telescope room, which were susceptible to interference from electrical noise were completely enclosed in solid copper enclosures. This included amplifiers, detectors (except for a small aperture to admit the signal), and co-axial cables (copper piping was used here). Nickel cadmium batteries were used to power all of the electronics shielded by these enclosures.

The signal from the LTT detector is amplified by a transresistance amplifier (labelled TC3) the output of which is connected to either an oscilloscope (used for detector alignment) or to the input unit of an analogue optic fibre link (Opto Tx). This unit transmits an optical pulse, linearly related in amplitude to the electrical input signal down a 20 m glass optic fibre to a receiver unit (Opto Rx) located in the "computer room". The optical pulse is then converted back into an electrical signal, again preserving the linearity between input and output pulse amplitudes, and digitised via the waveform recorder. Twelve microseconds prior to the LTT diode detecting a return signal from the topographic target (= 2 x 1800/c, see equation (3.9)), an

amplified signal from the CMT detector is summed at the input to the Opto Tx unit. This earlier signal represents a measurement of the transmitted power and is later used to normalise the return signal from the remote target. It is also used to trigger the waveform recorder.

A single "opto-electrical" channel is therefore used to monitor both the transmitted and received powers. Dual wavelength transmission consequently depends on an intra pulse separation of at least 12 μsec. An upper limit to the pulse pair separation is set by the sampling interval selected on the waveform recorder and the number of digital samples forming a complete record (up to 2048). The sampling interval must be sufficiently small to resolve the peak pulse amplitudes without excessive error. Components of the signal processing hardware are discussed in further detail below.

4.1.1 LTT Detector Characteristics

The responsivity of Lead Tin Telluride photovoltaic detectors to radiation of wavelength λ is defined as

$$R_I = \frac{\eta e \lambda}{hc} \quad AW^{-1} \tag{4.1}$$

where η is the quantum efficiency at wavelength, λ, e is the electronic charge, h is Planck's constant and c is the velocity of light. For the particular Plessey LTT diode used in these studies, the responsivity at 10.05 μm was given as 3.9 AW^{-1} and this is independent of frequency over the range 0 - 60 MHz [1]. The detector element, with an active area approximately 110 μm across, is mounted on a heat sink which is bonded to the inner wall of a liquid nitrogen dewar. This dewar has sufficient capacity for several hours operation. The entire assembly

is housed in an aluminium cylinder of 76 mm diameter by 205 mm.

Two sources of noise affect the performance of LTT diodes; Johnson Noise and Shot Noise. Johnson noise in the detector, which originates in resistive elements, can be expressed as

$$i_{jn} = 2\left[\frac{kT\Delta f}{R_o}\right]^{\frac{1}{2}} \qquad (4.2)$$

where k is Boltzmann's constant, T is absolute temperature, Δf is the noise bandwidth, and R_o is the detector slope resistance at zero bias. Shot noise is due to movement of discrete charge carriers across a potential barrier and is given by

$$i_{sn} = (2ei_p\Delta f)^{\frac{1}{2}} \qquad (4.3)$$

where e is the electronic charge and i_p is the induced photocurrent due to an optical power, P. Since $i_p = R_I P$, (4.3) may be re-written as

$$i_{sn} = (2eR_I P\Delta f)^{\frac{1}{2}} \qquad (4.4)$$

The incident power at which shot noise becomes equivalent to Johnson noise is therefore given by

$$P = \frac{2kT}{eR_I R_o} \qquad (4.5)$$

which for a slope resistance, R_o, of approximately 2kΩ [1], becomes ∿7 μW. It will be shown later (Section 4.2.1) that the incident powers actually detected fall in the range 0 - 15 μW, so that Johnson noise and shot noise will have comparable values.

4.1.2 The Transresistance Amplifier

The amplifier used (labelled TC3 in Figure 4.1) was based on

a design included with the Plessey application notes and used an SL561 video op-amp as the gain block [1]. This basic design was modified by Dr. S. E. Taylor who used it for a remote sensing system employing tunable laser diodes [2]. Circuit details and a frequency response diagram are available in reference [2] but, essentially, it exhibits a flat frequency response over the range 100 Hz to nearly 10 Mhz and has a transresistance gain, G_{TC3} of $5 \times 10^{4} VA^{-1}$.

For the purpose of the studies reported here, both the frequency response and linearity of the amplifier must be considered. Using the test circuit of Figure 4.2, the linearity was checked by applying a square wave signal of varying amplitude at the input denoted by V_{in}, and measuring the output voltage, V_{out}, generated by TC3. The input to the test circuit is terminated by a 47 Ω resistor and the simulated signal into the amplifier is given by

$$i_s = \frac{V_{in}}{R_s} \tag{4.6}$$

where R_s is 50 kΩ. The amplifier is assumed to have an input independence of less than 1 kΩ over the bandwidth indicated above [3]. To provide a test signal representative of the frequency content of the real signal, the input square wave actually consisted of a train of pulses, 350 nsec in width, with a separation of the order of milliseconds. The amplifier proved to be linear to within 5% over an input current range of 40 nA to 40 μA.

4.1.3 Optical Isolation Link

Early lidar system tests used a 20 m length of co-axial cable, completely sealed inside solid copper tubing, to relay the signal from the output of the amplifier to the waveform recorder in

an adjacent room. Despite numerous attempts at earthing, significant electrical interference was found to be present which corrupted the lidar returns and caused false triggering of the waveform recorder. Occasionally it would also cause the computer to "go down". As an answer to these problems an analogue optical communication channel was constructed which could transmit pulse height information from the transresistance amplifier to the waveform recorder. Considerable effort was expended to ensure that this channel was linear over the pulse height range of interest.

Figure 4.3 illustrates, in block diagram form, the configuration used to examine the linearity of the combined transresistance amplifier and optic fibre link. The total gain of this system is given by

$$G_{TOTAL} = \frac{V'_{out}}{V_{in}} = \frac{G_{TC3} G_{opto}}{R_s} \qquad (4.7)$$

The gain of the optical link was adjusted to unity, and since $R_s = G_{TC3}$ the output voltage should be identical to the input voltage. Figure 4.4 is a plot of the measured values of V'_{out} as a function V_{in}. As can be seen, although this system appears to be very linear, the actual value of G_{TOTAL} obtained is 1.9. This is assumed to be due to an impedance mis-match between the output of TC3 and the input of the optic fibre transmitter. A modified transresistance gain of $G = 1.9 \times 5 \times 10^4 = 9.5 \times 10^4 VA^{-1}$ is therefore assumed for the TC3-opto link combination. With the LTT detector replacing the test circuit the voltage expected at the output of the optic fibre receiver will therefore be

$$V'_{out} = Gi_P = G \frac{P}{R_I} \qquad (4.8)$$

Figure 4.5 is a circuit diagram of the optical transmitter and

receiver modules, both of which are designed around Plessey SL 541 op amps. The input to the opto transmitter is terminated with a 47 Ω resistor and is capacitively decoupled from the input to the op amp. An inverting amplifier configuration is used here having a gain defined by the resistors R3 and VR2. Transistor Q1 (BC 327) provides the driving current through (D3) and is biased to the on condition by the potential divider VR1. By adjusting VR1, and consequently the current through D3, the most linear region of the photodiode optical responsivity curve was empirically determined.

A pin diode is used as the input to the optical receiver and generates a voltage across R1, which is capacitively coupled to the non-inverting input of the Plessey op amp. The gain for this configuration is defined by VR1 and R4. A BC 373 npn transistor, biased to its linear region, is used in the output stage which is designed to drive a 50 Ω load. Both transmitter and receiver are powered by Nickel Cadmium batteries.

4.1.4 CMT Detector Monitoring of the Transmitted Power

From the lidar equation (3.7), derived in the previous chapter, it is apparent that a knowledge of the transmitted power, P_T, is required in order to determine the atmospheric absorption. It is an advantage of the differential absorption technique, however, that absolute values of the transmitted powers need not be measured. It will be demonstrated in Chapter 5 that, providing the component actually measured in the reference power channel is linearly related to the transmitted power, and that this linear relationship is identical for both wavelengths, then factors such as amplifier gain, attenuator and aperturing losses cancel out if additive measurement noise is assumed

to be negligible.

The Mercury Cadmium Telluride detector,(CWT), used for monitoring transmitted power was situated behind the 90% germanium beam splitter, as described in Chapter 2. Ten per cent of the output energy was therefore available for reference purposes. The important characteristics of the CMT detector are as follows:

Spectral Range	- 8 - 12 μm
Active Area	- 1 mm x 1 mm
Rise Time	- 2 nsec
Responsivity	- 1 - 50 mV/W
Maximum Signal	- 1 V
Bias Current	- 100 mA

A 5 cm (f/5) germanium lens was used to reduce the beam diameter to a dimension comparable with the detector active area and, to reduce the peak pulse power to below the damage threshold, polythene attenuators were inserted before the lens. The output from the CMT was coupled into the input of the optic fibre transmitter via a summing circuit, and was scaled so that the dynamic range of the reference power pulses matched that of the pulses detected by the LTT detector. A test was conducted to ensure the linearity of this arrangement by comparing the digitised peak pulse values monitored by the CMT detector, with an integrated power recorded using a Joule meter and a chart recorder. The Joule meter used the energy discarded by the 50% beam splitter in the laser room. Relative variations in the laser output were monitored as the laser voltage was adjusted, and both were found to agree to within 5%.

4.1.5 The Waveform Recorder

A "Biomation 8100 Transient Digitiser" was used to perform all analogue to digital conversions. This device has a specification which includes:

(i) Bandwidth - 25 MHz

(ii) Sampling Rate - 100 MHz (max.)

(iii) Resolution - 8 bits

(iv) Record Length - 2048 samples

(v) Read Speed - 10^6 Words/second

At a sampling interval of 10 nsec, therefore, a complete record would have a duration of approximately 20 μseconds. Lower sampling rates could be selected using decade adjustments over the range 0.01 μsec to 10 seconds. Sampling rates between these increments were also available, but only in decade multiples of 1, 2 or 5. A parallel I/O port was incorporated into the 8100 which allowed full software control of all the functions selectable from the front panel, and transfers of the 2048 word sample memory to the PDP 11 host computer. An interface had already been constructed for this purpose during earlier lidar studies [see ref. 4].

4.1.6 Data Capture and Software Control

The data acquisition program was written in Fortran but used an assembly language routine to read the 2048 samples from the transient digitiser into a Fortran array. Only four 8 bit words were required from each of these records; two bytes corresponding to the peak pulse heights of the two transmitted powers (measurement and reference wave-

lengths) and two bytes corresponding to the subsequent received powers. To minimise the amount of computer processing required (between shots) to identify these values, a windowing technique was used. Only those within an interval centred on the expected position of each pulse were examined. The width of this interval was approximately three times the FWHM of the pulse to allow for some jitter in the delay between laser shots.

All of the data analysed in Chapter 7 was captured at a sample interval of 50 nsec to permit the use of an intra pulse pair delay of 50 μsec. It was observed that jitter became more of a problem at smaller delays, and would necessitate an increase in the width of the data processing windows, with a consequent decrease in the dual wavelength prf. Sample intervals greater than 50 nsec would have unacceptably degraded the peak pulse height information because fewer samples would be taken over the pulse duration.

Figure 4.6 is a schematic of the software control and data acquisition system. It will be noticed that the interface to the PDP 11 also provides a trigger control for the firing of the lasers. Prior to the issuing of this command, the 8100 is re-initialised; this can only occur once all 2048 samples of the previous record have been clocked out and all four pulse heights have been identified. To eliminate any possibility of laser generated electrical noise stopping, or even damaging the computer, a digital optic fibre system was used where practical. The laser fire control consists of a single optical pulse which is intercepted by the laser controller. Intra pulse pair delays ranging from simultaneous to a few milliseconds can be set via the front panel of this device. Upon receipt of the command pulse, two optical pulses, with the pre-set delay, are transmitted to the Laser Trigger unit in

the laser room and are there converted into high voltage trigger pulses.

4.2 Direct Detection

4.2.1 Anticipated Signal Levels

An estimate of the signal levels expected can be formed using the lidar equation

$$P_R = P_T C \frac{\rho}{\pi} \frac{A_R}{r^2} e^{-2\alpha r} \qquad (3.7)$$

If the mean output energy from each of the lasers is $\sim$ 125 mJ and it is assumed that approximately one third of this is deposited in the gain switched spike, then the mean transmitted power would be

$$P_T = \frac{125 \times 10^{-3} \text{ mJ}}{3 \times 350 \text{ nsec}} = 1.2 \times 10^5 \text{ Watts}$$

An estimate of the transceiver efficiency, C, must include at least the following factors:

x 0.5 due to the 50% beam splitter losses (Tx path)

x 0.9 due to the 90% beam splitter losses (Tx path)

x 0.6 due to aperturing losses (see Figure 5.6)

x 0.9 due to the 90% beam splitter in the receiver path

giving an overall efficiency of $\sim$ 0.24. To allow for further, unknown losses, it will be assumed that C = 0.1.

Since a non-co-operative target was used, angled away from the beam path, it will be assumed that the Lambertian reflectivity, ρ, is also 0.1. Grouping terms together, it is found that

$$C \left(\frac{\rho}{\pi} \frac{A_R}{r^2} \right) = 2.8 \times 10^{-10}$$

Losses due to atmospheric absorption need to be estimated for at least two lines, corresponding to the reference and measurement wavelength for an absorbing species. Since H_2O is the most plentiful and absorbs at CO_2 wavelengths, a pair of recommended lines will be considered here: these are the 10 μm R branch transitions, R18 and R20, having absorption coefficients (α') of $0.950 \times 10^{-4} Atm^{-1} cm^{-1}$ and $0.851 \times 10^{-3} Atm^{-1} cm^{-1}$ (see Table 7.1).

If a relative humidity of 50% is assumed at a temperature of 15°C, then the concentration, γ, of H_2O is 8.4×10^{-3} yielding exponential terms of

$$e^{-2\alpha r} \text{ (for 10 R18)} = 0.75 \quad (2\alpha r = 0.287)$$

and

$$e^{-2\alpha r} \text{ (for 10 R20)} = 0.076 \quad (2\alpha r = 2.573)$$

The received powers at each wavelength are therefore estimated to be

$$P_R \text{ (10 R18)} = 25 \ \mu W$$

$$P_R \text{ (10 R20)} = 2.5 \ \mu W$$

Using (4.8), the voltages expected at the output of the optic fibre receiver (into a 50Ω load - the waveform recorder output) are therefore

$$V'_{out} \text{ (10 R18)} = 0.63 \text{ volts}$$

$$V'_{out} \text{ (10 R20)} = 0.063 \text{ volts}$$

The input voltage range actually selected for the waveform recorder was ± 0.2 V. An offset of x - 0.9 was used to give a "full scale deflection" of 0 to 0.39 volts, and this was found adequate for recording the full dynamic range of the received signals with only very occasional

clipping (<< 1 in 1000).

4.2.2 Contributions to Signal Statistics

Detector and Amplifier Noise

In Section 4.1.1 it was shown that Johnson noise and shot noise have similar values over the detected power range stated above. Shot noise was calculated there, however, by assuming that only the incident signal power was present. If the contribution due to the thermal background is now considered, the background radiation power on the detector may be calculated using [5]

$$P_b = A\Omega \frac{2h\nu^3 \Delta\nu}{c^2} e^{-h\nu/kT} \qquad (4.9)$$

where A is the detector area, Ω is the solid angle subtended by the detector beyond the regions cooled to 77 K ($\equiv$ 60° FOV), and $\Delta\nu$ is determined from the FWHM of the detector responsivity curve [obtained from ref 1]. This gives a value for P_b of 7.9×10^{-7} W.

The various noise currents may now be determined using expressions (4.2) and (4.4):

$$i_{jn} = 9.1 \text{ nA}$$

$$i_{sn} (P_s = 7.5 \text{ μW}) = 9.7 \text{ nA}$$

$$i_{sn} (P_b = 0.8 \text{ μW}) = 3.1 \text{ nA}$$

so that

$$i_{total} = i_{jn} + i_{sn}(P_s) + i_{sn}(P_b) \approx 22 \text{ nA}$$

where the mean incident signal power is assumed to be 7.5 μW, and a

bandwidth $\Delta f = 10^7$ Hz is assumed to apply both the transresistance amplifier and the optic fibre link. Applying, once more, the gain relationship of (4.8), the noise voltage expected at the output of the fibre optic receiver is therefore ~ 2.6 mV. This is approximately two orders of magnitude less than the signal voltage generated by a mean incident power of 7.5 μW.

No attempt is made here to calculate amplifier noise but a noise measurement was obtained by adjusting the input voltage range of the waveform recorder down to ±0.05 V, where noise caused continuous triggering. Observed peak to peak variations were confined to no more than ± 10 mV and this would, of course, include both detector and amplifier noise in the absence of signal.

Digitisation Noise

Two forms of digitisation noise will affect the measured peak pulse voltages:

(i) Quantisation noise:

This is a consequence of the finite number of digital levels allocated to a given voltage range. If V represents the full scale deflection and n is the number of binary digits used, the quantisation error may be expressed as

$$\sigma^2_{lin} = \frac{(\Delta V)^2}{12} \qquad (4.10)$$

where

$$\Delta V = \frac{V}{2^n - 1} \qquad (4.11)$$

[see, for example, ref. 6]. For the Biomation 8100, n = 8, so if V is assumed to be 0.2 V (½ fsd), the voltage error is

$$\sigma_{lin}(V) = 0.23 \text{ mV}$$

(ii) Sampling noise:

This is a consequence of the finite number of samples taken over the pulse duration. For a pulse of duration τ_p (FWHM) and a sampling period, τ_s, the average number of samples occurring during any interval, τ_p, will be

$$N_s = \frac{\tau_p}{\tau_s} \tag{4.12}$$

which has a value of 7 for the measurements considered here. Only 50% of these on average will occur between the half maximum point, and the peak, on either side of the pulse. An approximate expression for the normalised voltage error in this case is given by

$$\frac{\Delta V}{V} = \frac{1}{N_s} \tag{4.13}$$

which is equal to 0.14 or 14% for $N_s = 7$. In voltage terms this corresponds to nearly 30 mV for V = 0.2 volts. Although this is relatively large, it will be shown below that speckle noise forms the dominant noise source. Sampling errors are, in any case, not a fundamental problem if the latest waveform recorders or digital oscilloscopes are used.

Speckle

The surface of the topographic target, in common with the vast majority of surfaces, is extremely rough on the scale of an optical wavelength. When illuminated by a coherent light source, the optical field reflected from such a surface consists of contributions from many

independent scattering areas [6]. At a remote observation point, these various scattered components will have relative delays which may vary from several to many wavelengths, depending on the surface roughness and the propagation geometry. Their interference at the observation point results in the granular pattern known as speckle, consisting of "bright spots" where the interference has been highly constructive, dark region where it is highly destructive, and irradiance levels between these extremes.

If the observation point is moved, the path length travelled by the scattered components changes and a new speckle pattern observed. Similarly, if the position of the illuminated spot on the rough surface is moved a new, statistically independent speckle pattern is generated at the observation point providing the distance moved is at least $\lambda/2$. The contributions from elementary scatters may be described in terms of a phasor having real and imaginary components in the complex plane. The amplitude a_k and phase ϕ_k of the kth elementary phasor are statistically independent of each other, and also of the amplitude and phases of all elementary phasors [6].

Providing the number of these elementary contributions is large, their resultant in the complex plane has real and imaginary parts which are independent, zero mean, identically distributed Gaussian random variables, with a phasor angle ϕ, which is uniformly distributed over the interval $-\pi$ to $+\pi$ (see Figure 4.7). At any given point (x,y,z) the field amplitude is found to be Rayleigh distributed but the intensity has an exponential distribution described by

$$P(I) = \left[\frac{1}{\bar{I}}\right] \exp \left[-\frac{I}{\bar{I}}\right] \qquad (4.14)$$

where $\bar{I}$ is the mean or expected irradiance. A fundamental property of

this distribution is that its standard deviation precisely equals its mean. Goodman points out that the contrast of a polarised speckle pattern, as defined by

$$C = \sigma_I / \bar{I} \tag{4.15}$$

is always unity; hence the subjective impression (for visible speckle) that the irradiance fluctuations are a significant fraction of the mean [6].

At the remote (far field) observation point (x,y,z), the approximate size, or lateral coherence distance, of a single speckle lobe is [6]

$$d_s = \frac{\lambda L}{D_s} \tag{4.16}$$

where D_s is the diameter of the illuminated spot, λ is the wavelength of the radiation and L is the distance to the observation point (assumed perpendicular to the scattering plane). If an imaging system is employed, the speckle pattern is generated by the interference of light from two points on the lens aperture, giving a lateral coherence distance of

$$d_s(\text{lens}) = \frac{\lambda z}{D_L} \tag{4.17}$$

where z is the distance from the lens to the image plane and D_L is the lens diameter. Consequently, for the lidar system considered here, ($z = 1.8$ m, $D_L = 0.3$ m), $d_s(\text{lens}) = d_s(\text{receiver mirror}) = 60$ μm. In other words, the width of each speckle is nearly half the dimension of the detector active area.

The speckle pattern at a fixed point will change with time if the position of the spot on the target moves. If the scatterers on one side of the target spot move with respect to those on the opposite side by a distance of $\lambda/2$, the addition of the scattered fields at the receiver

produces an entirely independent power sample. It is possible to show [5] that the characteristic time constant of these changes, τ_s, may then be expressed as

$$\tau_s = \frac{\lambda}{2D_s V} \tag{4.18}$$

where V is the velocity of the spot across the target. Consequently, for the power samples at the receiver to decorrelate over a period of $\sim$ 1 second, the velocity of the target spot needs to be only

$$V = \frac{\lambda}{2D\tau_s} = 2.5 \times 10^{-5} ms^{-1} \tag{4.19}$$

assuming that D_s = 0.2 m.

This result is surprising since correlation in CO_2 lidar data is usually observed over approximately this period of time, and is usually attributed to changes in the state of the atmospheric channel (see Chapter 5 and 7). Also the correlation coefficient between wavelength channels, for $\lambda_1 = \lambda_2$ was observed to the $\sim$ 0.9 for the lidar system described here, suggesting near perfect alignment of the beam paths. A possible solution is that the target surface geometry actually possesses a degree of spatial coherence over distances much larger than $\lambda/2$. Atmospheric turbulence, however, may be expected to increase the independence of power samples since it will result in higher spatial frequencies in the distribution of energy at the target, both within and between laser pulses.

The exponential power distribution of (4.14) will apply whenever the lateral coherence distance of the imaged speckle pattern is greater than or equal to the detector dimension. For the receiver configuration considered here, however, the number of speckle lobes imaged at the detector surface will be approximately 3 or 4 (d_s = 60 μm, detector

dimension = 110 μm). Another probability density function is therefore required to describe this more general situation. In Chapter 5, the Chi-square distribution function is introduced as an appropriate model, applicable when one or more speckles are imaged at the detector (see expression (5.19)). A parameter, m, which may be identified with the average speckle count, is used to control the shape, mean and variance of the distribution.

Of primary importance here is the ratio of the standard deviation to the mean, given by $1/\sqrt{m}$, which provides a direct indication of the magnitude of speckle induced noise. For m = 1 (the exponential power distribution), speckle noise is at its worst since variations in the signal are comparable to its mean value. A speckle count of m = 4, as applies here, yields a normalised standard deviation of 0.5. Although less than for m = 1, speckle induced noise is still much larger than the other noise sources considered in previous sections. The data processing techniques described in Chapter 6, and applied in Chapter 7, reflect the need to suppress the degrading influence of speckle.

REFERENCES

[1] "Lead Tin Telluride Photovoltaic Detectors", Application Notes, Plessey Optoelectronics and Microwave Ltd., P54785, Issue 2.

[2] TAYLOR, S.E. "Tunable Diode Laser Instrumentation for Remote Sensing", Ph.D. Thesis, Dept. of Applied Physics, Hull University, England, 1984.

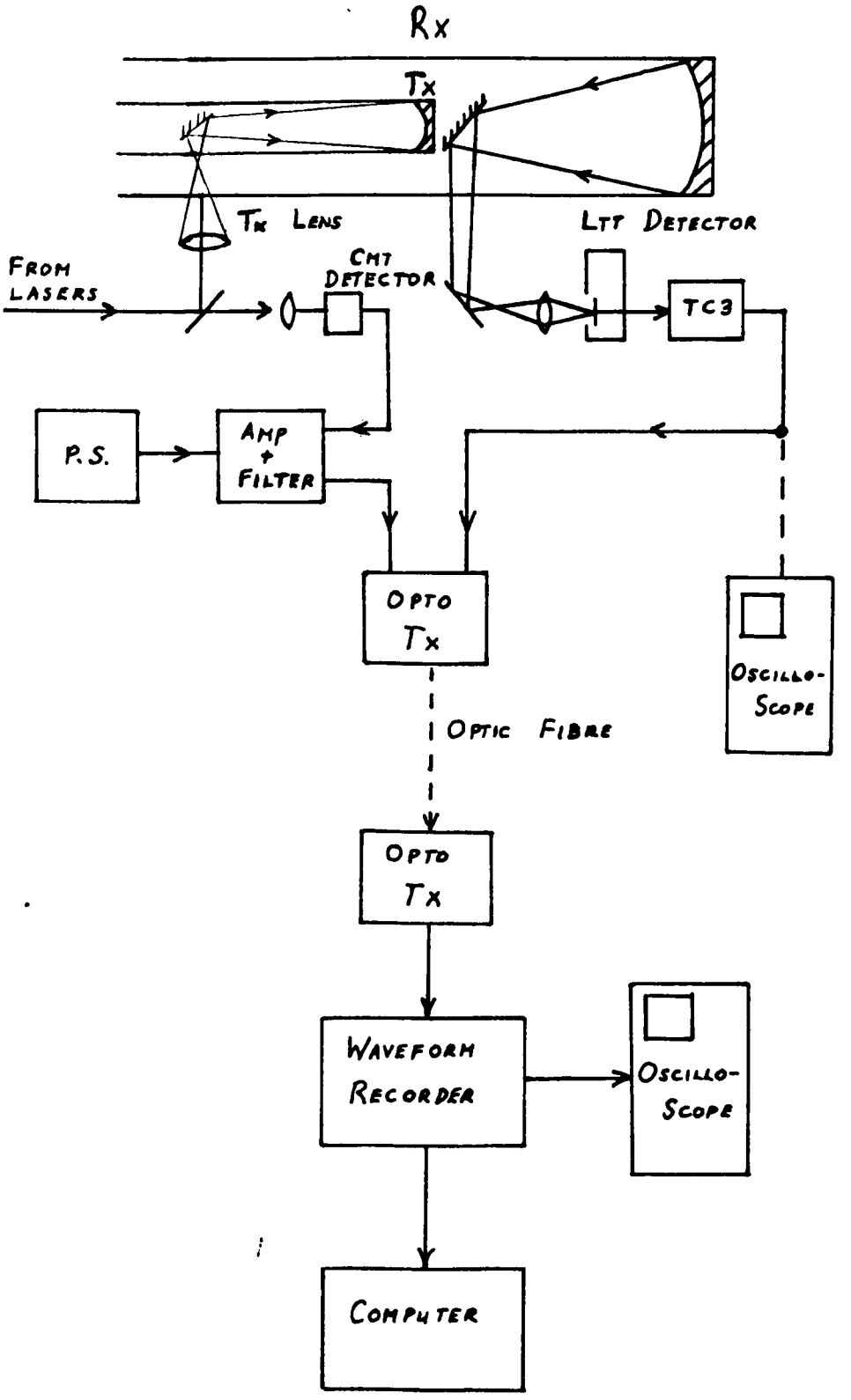
[3] TAYLOR, S.E. Private Communication.

[4] GREEN, B.A. "A Coherent Laser Radar for Atmospheric Pollution Monitoring", Ph.D. Thesis, Dept. of Applied Physics, Hull University, 1979.

[5] KINGSTON, R.H. "Detection of Optical and Infra-red Radiation", Springer-Verlag, 1978.

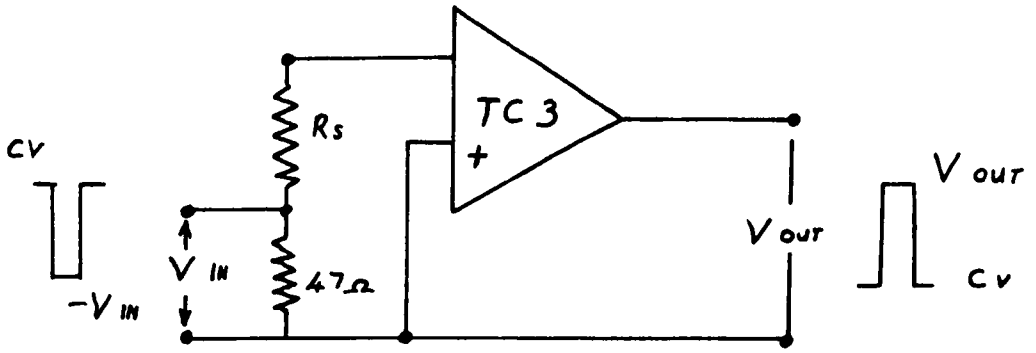
[6] BETTS, J.A. "Signal Processing, Modulation and Noise", Hodder and Stoughton, 1981, ISBN 0 340 09895 e.

[7] GOODMAN, J.W. "Some Fundamental Properties of Speckle", J. Opt. Soc. Am. 66(11), Nov. 1976, pp. 1145-1150.



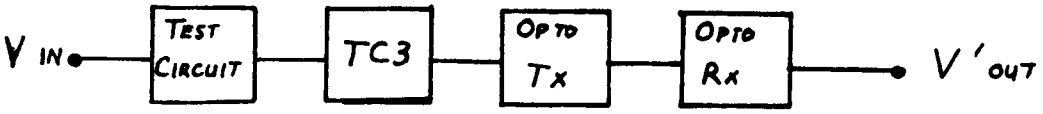
Tranceiver and Signal Processing Components

Fig. 4.1



Signal Simulation Circuit at input to TC 3 Amplifier

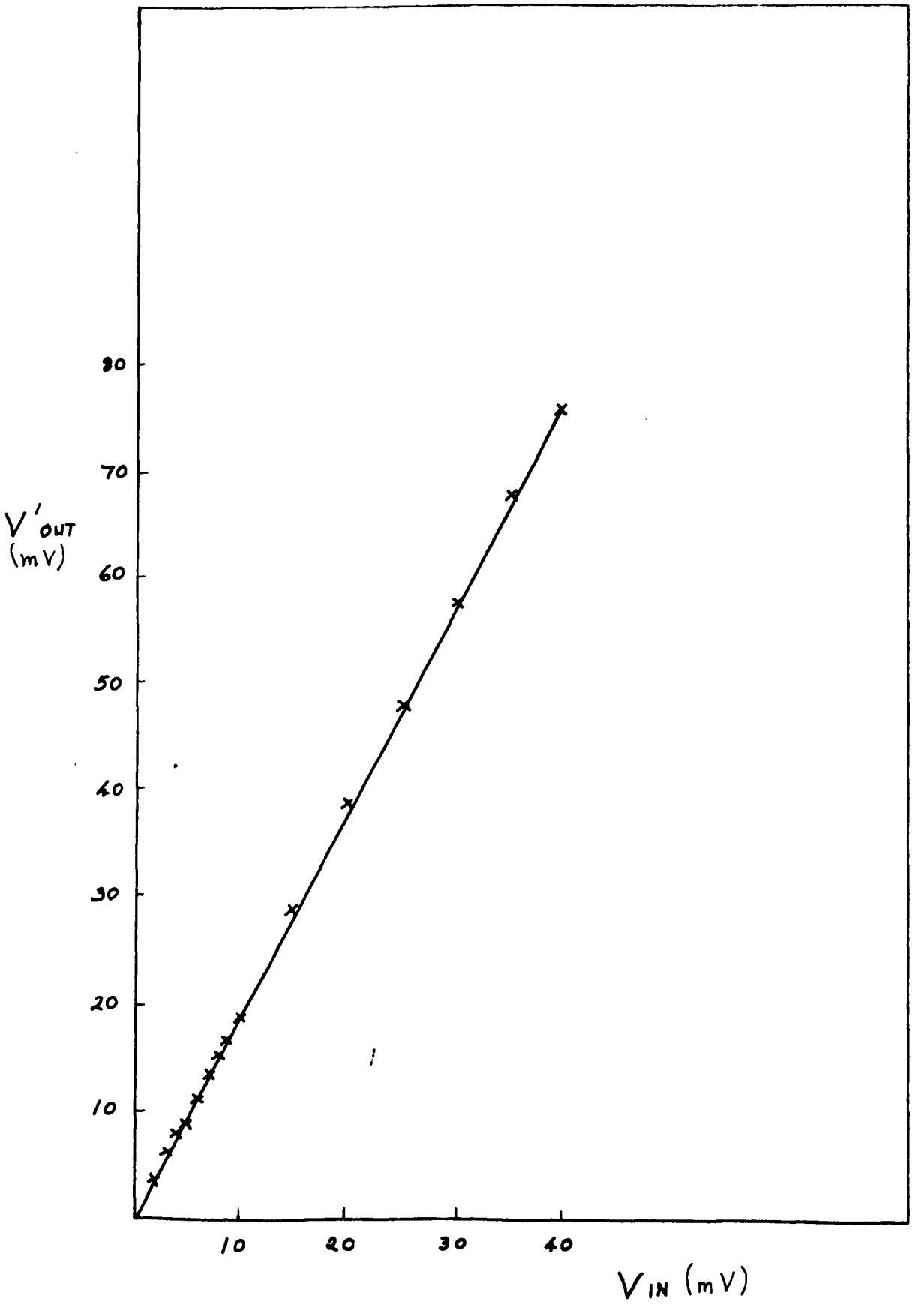
Fig. 4.2

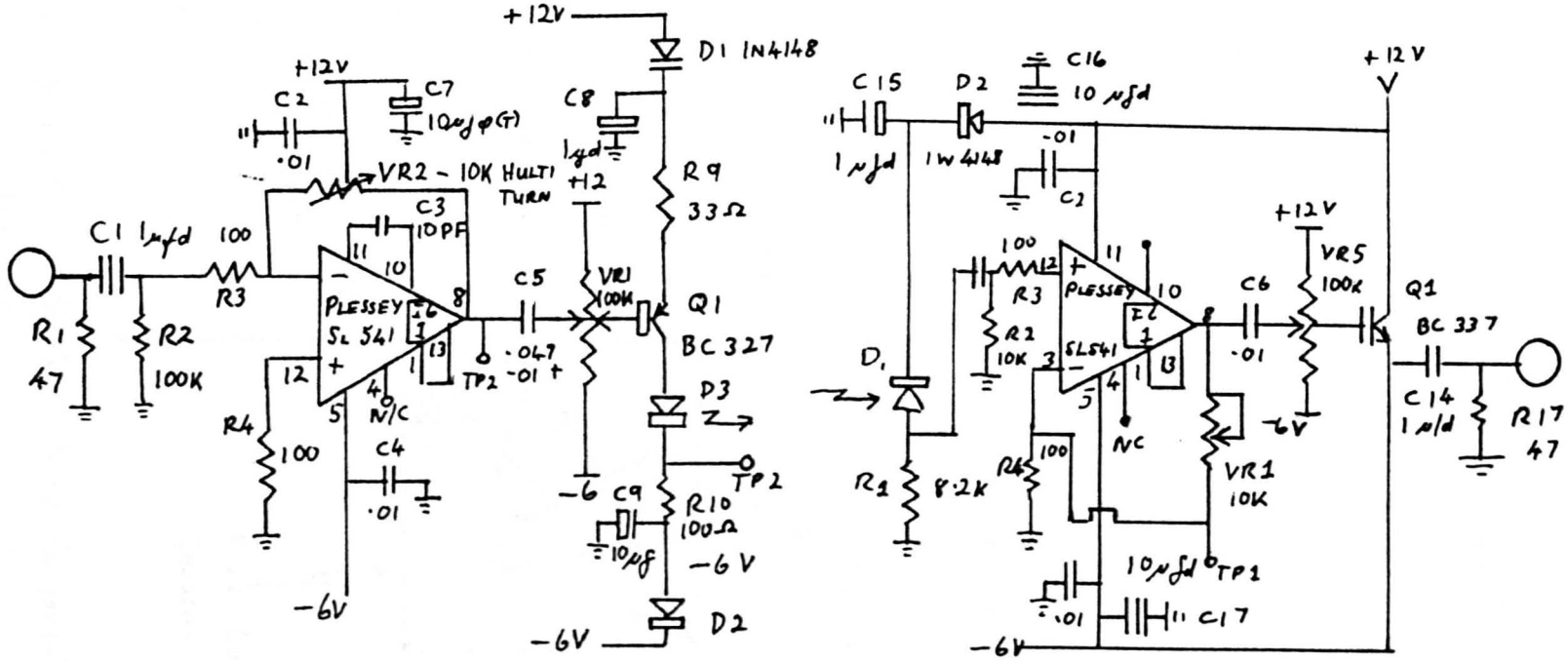


Signal Processing Module Linearity Test Configuration

Fig. 4.3

Fig. 4.4

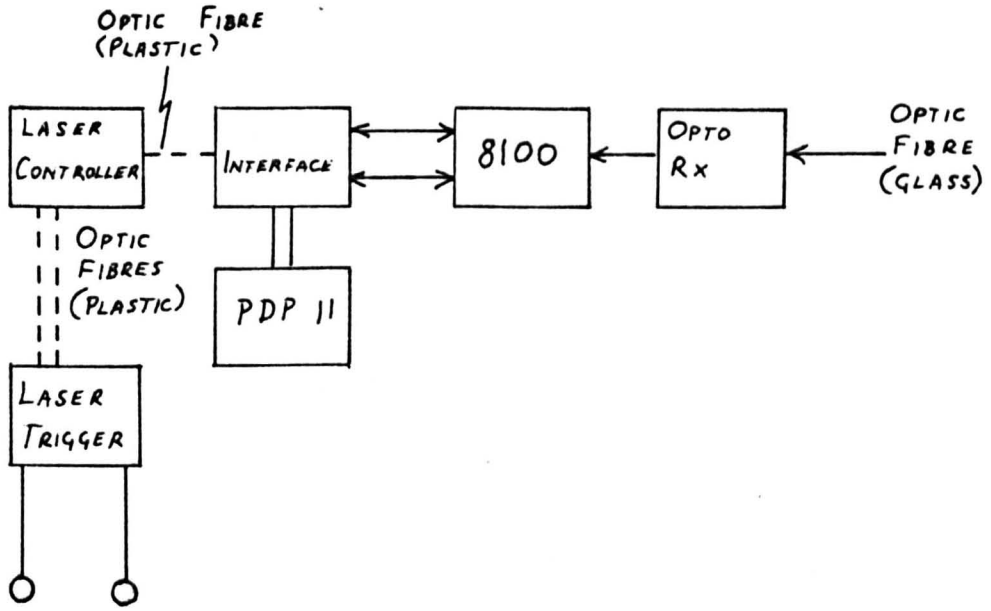




OPTO - TRANSMITTER

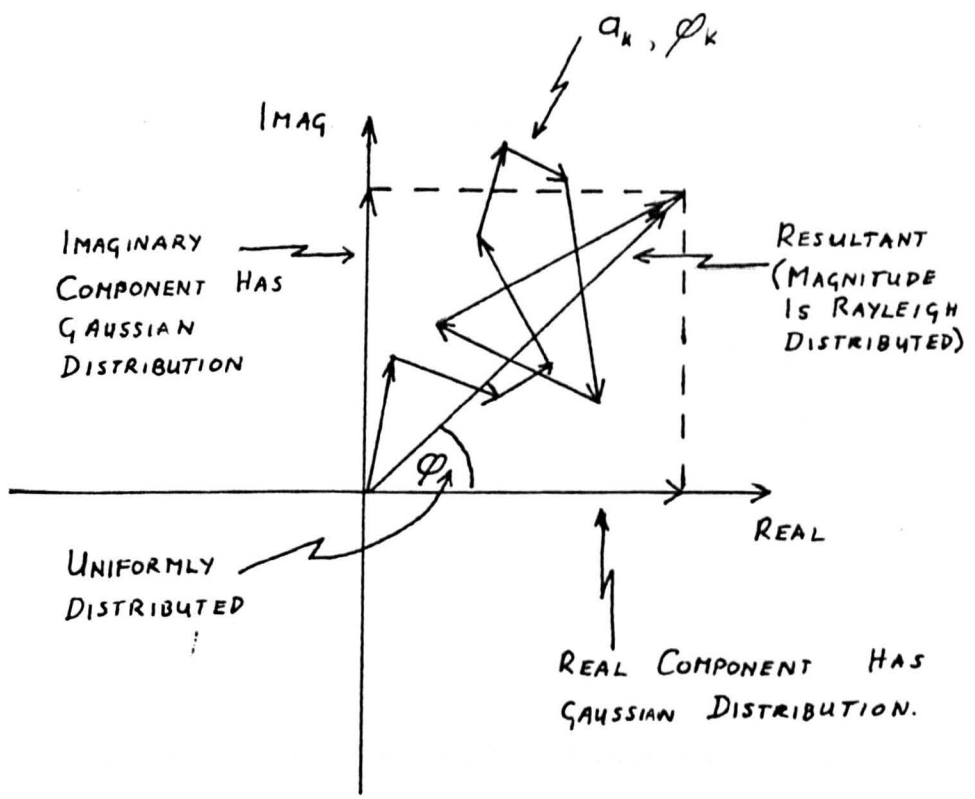
OPTO - RECEIVER

Fig. 4.5



Optic Fibre Usage in Laser Control and Signal Processing

Fig. 4.6



Statistical Distribution of Speckle - Phasor Components

Fig. 4.7

CHAPTER 5.0

MEASUREMENT SENSITIVITY

The precision with which the concentration of a particular molecular species may be measured depends on the combined uncertainty associated with each of the parameters included in equation (3.29). It is possible to express this combined uncertainty as [1,7]

$$\frac{\sigma^2(\gamma)}{\gamma_m^2} = \frac{\sigma^2(K)}{K^2} + \frac{1}{4K^2r^2n^2}\left\{\frac{\sigma^2(P_{N1})}{\bar{P}_{N1}^2} + \frac{\sigma^2(P_{N2})}{\bar{P}_{N2}^2} - \frac{2\,\mathrm{Cov}[P_{N1},P_{N2}]}{\bar{P}_{N1}\,\bar{P}_{N2}}\right.$$

$$+ \frac{\sigma^2(C_1)}{\bar{C}_1^{\,2}} + \frac{\sigma^2(C_2)}{\bar{C}_2^{\,2}} - \frac{2\,\mathrm{Cov}[C_1,C_2]}{\bar{C}_1\,\bar{C}_2}$$

$$\pm \frac{2\,\mathrm{Cov}[P_{N1},C_1]}{\bar{P}_{N1}\,\bar{C}_1} \pm \frac{2\,\mathrm{Cov}[P_{N2},C_2]}{\bar{P}_{N2}\,\bar{C}_2}$$

$$\left.\pm \frac{2\,\mathrm{Cov}[P_{N1},C_2]}{\bar{P}_{N1}\,\bar{C}_2} \pm \frac{2\,\mathrm{Cov}[P_{N2},C_1]}{\bar{P}_{N2}\,\bar{C}_1}\right\} \quad \ldots \quad (5.1)$$

where the target range, r, and the number density of molecules in the atmosphere, n, are assumed to be known precisely. P_N denotes the

normalised power ratio P_R/P_T so that P_{N1} and P_{N2} correspond to the normalised power returns at wavelengths 1 and 2 respectively. The various terms identified above as contributing to inaccuracies will now be examined.

5.1 Uncertainties in the Absorption Coefficient

5.1.1 Uncertainties in α: Interfering Species

The successful application of line tunable CO_2 laser DIAL to pollutant and trace gas concentration measurements depends on a range of criteria being satisfied with respect to coincidences in the CO_2 laser line emission spectra and the absorption spectra of the gas to be monitored. These may be summarised as follows:

(1) A suitable line in the CO_2 laser line emission spectra must overlap a suitable absorption feature in the spectra of the pollutant or trace gas. Since such overlaps are simply a matter of chance it is extremely unlikely that the line centre frequencies, ν_o of the emission and absorption lines will coincide. The absorption term due to this overlap will therefore result from the convolution of two lineshape functions with comparable line widths.

(2) Having established a coincidence of the type defined above, a further requirement in that no other absorption feature overlaps the selected CO_2 laser line. This is equivalent to restricting the aggregate expressed in equation (3.21) to a single absorber and, in practice, is usually impossible to achieve using line tunable lasers. It has been suggested that if the background absorption is more than about 10% of the absorption due to the

gas species being monitored, then an accurate determination of pollutant concentration is probably impossible [2].
Interference by other absorbing species will include the contributions due to features such as the "water vapour continuum" and particles such as aerosols nucleated on solid matter (dust, soot, etc.) which exhibit a distribution of particle sizes depending on such factors as geographical location of the air mass, humidity and temperature [2]. Aerosol absorption is likely to be very difficult to predict but other sources of interference, such as that due to naturally occurring CO_2, may be estimated more accurately.

(3) Having satisfied the requirements of (1) and (2) for the measurement wavelength, a further requirement of the differential absorption technique is the selection of a suitable reference wavelength. A related set of constraints are imposed upon the selection of this wavelength:

(a) It is important that it be attenuated far less by the molecular species being examined, than the measurement wavelength is itself. This is necessary to maximise the differential absorption coefficient (DAC) since the magnitude of this quantity influences directly the fractional uncertainty contributed to the final concentration measurement by any uncertainty in the DAC (see expression 5.1).

(b) Any interfering species should ideally have identical absorption coefficients at both the measurement and reference wavelengths so that the measured differential

absorption is due solely to the molecule which is the subject of the measurement. This is clearly unlikely, since absorption features at atmospheric pressure usually have similar line widths and, any wavelength separation sufficient to generate a significant differential absorption term for one absorber, is likely to do the same for another.

(c) Conflicting, to a certain extent, with the requirement to maximise the DAC, is the additional requirement to maintain relatively close frequency proximity between the measurement and reference wavelengths so that differential scattering terms do not contribute as differences in the return powers. This is likely to become significant only when the DAC of the target molecule is relatively small and the separation between measurements and reference lines relatively large.

Many experiments have been conducted to determine the absorption coefficients of various atmospheric molecules and pollutants as a function of wavelength at various partial pressures and temperatures using a range of different radiation sources. Much of this information has been compiled into a database form and is available on magnetic tape ready for computer processing. An important example is the AFGL database tapes [3] which list parameter values such as the line strength at a particular wavelength, the line width, and a series of values corresponding to the rotational and vibrational parameters, identified in equations (3.16) to (3.29), for a collection of 14 different molecules.

Petheram has used this data, together with empirically derived models, for estimating both absorption due to the water vapour continuum,

and the scattering and absorption due to various aerosol distributions, to analyse the suitability of various pairs of CO_2 laser lines for DIAL applications [2]. Many of these lines have been recommended before but often without reference to the various sources of absorption interferences which would degrade their performance. Petheram's report investigates their potential impact on DIAL accuracies as a function of the concentration of interfering species, particularly water vapour, and includes various observations on temperature dependencies.

Species interference is therefore acknowledged as an important limitation to the ultimate performance of DIAL systems but is not pursued here since separate investigations are currently addressing the problem [4,5,6,10]. This work is restricted to identifying optimum techniques for reducing the variance in the log power ratio terms of equation (3.29).

5.1.2 Uncertainties in K(ν)

The differential absorption coefficient for a single molecular species was identified earlier as

$$K(\nu) = K_m(\nu_1) - K_m(\nu_2)$$

and will contribute an uncertainty term

$$\sigma^2(K) = \sigma^2(K_m(\nu_2)) + \sigma^2(K_m(\nu_1)) \qquad (5.2)$$

where the subscript 1 indicates the reference wavelength, and subscript 2, the measurement wavelength. To maximise the differential absorption coefficient ν_1 would have to be located away from the absorption peak selected for the measurement and between other peaks in the absorption

spectra of the target molecule. Even in the absence of interference absorption spectra, the reference wavelength absorption is therefore likely to be non-zero.

If (3.19) is re-written as,

$$K_m(\nu) = \frac{S\,\Delta\,\nu_\rho}{[\pi\Delta\nu^2 + \Delta_\rho^2]} \tag{5.3}$$

where $\Delta\nu = \nu - \nu_o$, the contributions to the uncertainty of K can be identified by expanding the above expression to resolve component variances so that,

$$\frac{\sigma^2(K_m)}{E[K_m]^2} = \frac{\sigma^2(S)}{E[S]^2} + \frac{\sigma^2(\Delta\nu_\rho)}{E[\Delta\nu_\rho]^2} + (-1)^2\,\frac{\sigma^2([])}{E[[]]^2} \tag{5.4}$$

where [] denotes the terms collected in the square brackets of (5.3). Taking the expansion a stage further,

$$\frac{\sigma^2(K_m)}{E[K_m]^2} = \frac{\sigma^2(S)}{E[S]^2} + \frac{\sigma^2(\Delta\nu_\rho)}{E[\Delta\nu_\rho]^2} + \ldots \tag{5.5}$$

$$+ \frac{4}{E[\pi\Delta\nu^2 + \nu_\rho]^2}\left[\frac{\pi^2[\sigma^2(\nu) + \sigma^2(\nu_o)]}{E[\Delta\nu]^2} + \frac{\sigma^2(\Delta\nu_\rho)}{E[\Delta\nu_\rho]^2}\right]$$

variances due to all the terms of (5.3) appear but, upon substitution of the expression (3.18) and (3.20), yet further expansion clearly becomes possible.

However, it is not necessary to pursue further expansion here to realise that ultimately uncertainties in temperature and pressure will contribute significantly to the overall uncertainty present in the differential absorption coefficient. Although experimental errors may influence, for example, the form of the quoted line intensity partition

function of (3.18), and quoted reference line widths, $\Delta\nu\rho_o$ of (3.20), these are potentially rectifiable via further calibration using laboratory samples. As such they will appear as systematic errors.

Errors due to unknown temperatures and pressures, however, are likely to be more significant [7]. If temperature and pressure can be regarded as constant throughout the measurement (approximately 20 minutes) they must be independently estimated for accurate concentration measurements, since averaging of pulse pairs will not reduce the measurement error due to discrepancies between actual and assumed values [7]. On the other hand, significant temporal and spatial variations in the magnitude of temperature and pressure during the course of the measurement may be encountered if, for example, the lidar beam intersects with several atmospheric convection cells. A certain degree of spatial and temporal averaging would result, but it would be extremely difficult to estimate accurately the effective mean and variance during the measurement, except, perhaps, by resorting to yet another lidar system specifically designed for the purpose, or by including extra reference wavelengths in a single lidar system to assist in eliminating unknowns.

5.1.2.1 Longitudinal Mode Energy Distribution

Hardesty identifies, as a further source of absorption coefficient error, the inexact knowledge of the distribution of transmitted energy among the various possible longitudinal laser cavity modes [7]. He points out that the absorption coefficient $K_m(\nu)$, is actually a weighted average of the form

$$K_{ave} = \frac{K_m(\nu_1)\,\varepsilon_1 + K_m(\nu_2)\,\varepsilon_2 + \ldots + K_m(\nu_n)\,\varepsilon_n}{\sum_{i=1}^{n} \varepsilon_i} \qquad (5.6)$$

where ε_i is the energy deposited into a longitudinal mode of frequency ν_i. K_{ave} may therefore be expected to vary on a shot to shot basis as the energy is redistributed and each ε_i changes in magnitude.

The Hull lidar system uses a cavity length of approximately 2.4 metres and since longitudinal cavity modes are spaced at frequency intervals of C/2L, where C is the speed of light and L the cavity length, a separation of 60 MHz between longitudinal cavity modes can be expected. Each CO_2 emission line has a width of approximately 3 GHz at atmospheric pressure and would therefore provide gain for up to $3 \times 10^9/60 \times 10^6$ = 50 longitudinal laser modes overlapping with this line width interval.

However, both CO_2 laser cavities included not only a pulsed TEA gain section but also a low pressure CW section to provide preferential gain to only those modes corresponding in frequency location to the $\sim$100 MHz linewidth characteristics of CW CO_2 sections operating at a pressure of 15 torr. Therefore, a maximum of two longitudinal modes would have existed during these experiments and without knowledge of how the total pulse energy was distributed between these modes, an uncertainty in the absorption coefficient, $K_m(\nu)$, corresponding to an uncertainty of $\sim$ $\pm$ 50 MHz in the value of ν, must be assumed.

5.1.2.2 Laser Stability

Although the line centre frequency of both CO_2 emission spectra and atmospheric molecular absorption spectra are fixed in frequency space, both the absolute frequency of longitudinal laser modes and their frequency separation will vary as the laser cavity length alters (due primarily to long term temperature trends within the laser cavity structure over the course of the measurement). Therefore,

not only does the fraction of the total transmitted pulse energy, ε_i, deposited per mode alter with time, but so does the frequency location of each mode, ν_i, relative to the line centre of emission and absorption spectra.

5.2 Logarithmic Terms

In order to form an estimate of $\log_e (P_N(\lambda_1)/P_N(\lambda_2))$, as required in expression 3.29, it is necessary to use a system of detectors and amplifiers to measure and normalise the return power with the result that the quantities actually ratioed are discrete voltage levels linearly related to the received powers.

Signal processing components identified in the previous chapter are reproduced in the schematic of Figure 5.1.

Speckle noise is regarded here as measurement noise and is multiplicative in nature since its variance changes linearly with changes in the mean of the return power. Therefore, if $P_R(k)$ is regarded as the integrated optical power incident on the detector at the kth shot and $\bar{P}_R$ is the mean return power at the kth shot in the absence of speckle, then

$$P_R(k) = \bar{P}_R(k)\, S(k) \tag{5.7}$$

where S(k) is the speckle noise.

S(k) can be regarded essentially as a random variable with a mean of unity ($\bar{S}(k) = 1$) such that,

$$E[P_R(k)] = \bar{P}_R(k) \tag{5.8}$$

Figure 5.2(a) is a histogram representation of the normalised return

power data for 10,000 shots which may be compared with a simulated data file (Figure 5.2(b)), also containing the equivalent of 10,000 shots, for which a Chi-Square random sequence of order 3 was used to generate the speckle noise (see section 5.2.3). Depending upon the temporal separation of the laser returns, it may be necessary to complicate this speckle noise model further by including temporal correlation of the speckle noise sequence.

The digitised voltage level corresponding to the peak power of the return pulse (see Chapter 4) is denoted by V_R and is related to the power incident on the detector by

$$V_R = P_R G_R + N_R \tag{5.9}$$

where G_R is the cascaded gain (volts per watt) of all amplifying components prior to digitisation and N_R represents the total of all noise sources including background radiation noise, amplifier noise and digitisation noise in the received power channel. A similar expression exists for the reference component of the transmitted power

$$V_T = P_T' G_T + N_T \tag{5.10}$$

so that the quantity used to normalise the return power is

$$V_N = \frac{P_R G_R + N_R}{P_T' G_T + N_T} \tag{5.11}$$

The pulse index, k, has been omitted from equations (5.9) to (5.11) for reasons of clarity and will also be omitted from subsequent expressions.

If P_R is now replaced by expression (5.7), and P_T' by $P_T F$, where F represents that fraction of the transmitted power extracted for refer-

ence purposes, then

$$V_N = \frac{\bar{P}_R.S.G_R + N_R}{P_T.F.G_T + N_T} \tag{5.12}$$

For every pulse pair corresponding to the measurement and reference wavelengths, there exists a corresponding pair of values $V_N(\lambda_1)$ and $V_N(\lambda_2)$ which are used to form the log ratio estimator.

$$\log_e\left[\frac{V_N(\lambda_1)}{V_N(\lambda_2)}\right] = \log_e\left\{\frac{\left[\frac{\bar{P}_R(\lambda_1).S(\lambda_1).G_R + N_R(\lambda_1)}{P_T(\lambda_1).F.G_T + N_T(\lambda_1)}\right]}{\left[\frac{\bar{P}_R(\lambda_2).S(\lambda_2).G_R + N_R(\lambda_2)}{P_T(\lambda_2).F.G_T + N_T(\lambda_2)}\right]}\right\} \tag{5.13}$$

5.2.1 Bias

In principle, possible estimators for the log ratio can be formed by either

(a) ratioing the powers and then forming the log of the ratio, or

(b) logging the powers and then forming the differences of the logs.

Estimators formed simply from the power ratios, which assume that the log operation is performed on the estimator, are not considered in this work because of the difficulties of estimating a quantity in the presence of multiplicative (speckle) noise.

Both (a) and (b) would represent identical operations on a single pair of power values but, when forming an estimate from an ensemble of normalised power values, the use of either of these transformations will result in bias. If the quantity to be estimated is stationary ($\bar{P}_R$ = a constant), a correction term can be applied, but not otherwise. In each case bias is attributable to [8]:

(i) The non-linearity of the ratio operation; speckle noise in both the measurement and the reference channels causes normalised power returns to fluctuate between zero and some multiple of the mode of the distribution as indicated in Figure 5.2. Forming the ratio of two such statistically varying quantities is clearly a non-linear operation since a singularity is approached as denominator values approach zero.

(ii) The non-linearity of the log operation; speckle noise is assumed to have a mean of unity but, regardless of the probability distribution assumed for the speckle sequence, the mean of such a series, under the log transformation, would only approach zero as the variance of the series approached zero. It is worth noting here that, as a consequence of this non-linearity, expression (5.1) is only an approximation. As the normalised power variance increases, high (even) order moments become important in the series expansion used to express the variance of the logarithm of a variable in terms of the variance of the variable itself [1].

(iii) The need to reject negative power data acquired in the presence of additive measurement noise; referring again to Figure 5.2. it is clear that speckle noise represents instantaneous fluctuations in the power falling on the detector and therefore cannot be less than zero. However, the additive noise terms in equation (5.13) can cause the voltage ratio in this expression to drop below zero and, since the logarithms of negative values do not exist, these values must be discarded.

If the noise is entirely multiplicative (i.e. due to speckle only) and positive definite, the ratio estimator is biased [9,10]. On the other hand the log ratio estimator is unbiased both in this limit and in the presence of additive noise when the ratio of the signal to additive noise ratios (SNRs) in the two channels is unity. This may be expressed in the form

$$\frac{\left[\dfrac{\text{Signal Voltage}}{\text{Additive Noise Voltage (Non-Speckle)}}\right]_{\lambda=1}}{\left[\dfrac{\text{Signal Voltage}}{\text{Additive Noise Voltage (Non-Speckle)}}\right]_{\lambda=2}} = \rho \qquad (5.14)$$

where $\rho = 1$.

Computed bias data is presented in Figure 5.3 for the case of analogue signals in uncorrelated Gaussian noise [10]. This shows the fractional bias resulting from the use of ratio and log-ratio estimates in the presence of uncorrelated additive noise (no speckle) as a function of the voltage SNR in the ratio numerator (abscissa) and the ratio of the voltage SNRs in the numerator and denominator (ρ). If the absorption fluctuates, the bias has to be weighted over the whole range of ratio and SNR values encountered.

5.2.2 Systematic Errors

Figure 5.4(a) shows typical transmitted and received powers, in both channels, as recorded in digitised form by the transient waveform recorder (sample time is 50 n sec and pulse widths are 350 n sec

FWHM). Between pulses, the noise levels are clearly very low compared to typical pulse levels. A similar comparison can be made over the duration of a typical measurement sequence using an estimate of the mean level of the received power. Examples obtained using a suboptimal recursive filter are presented in Figures 5.4(b) and 5.4(c) for the transmitted and received power respectively. Digital units along the ordinates of both figures correspond to the same power levels, and values along the abscissae correspond to laser shot numbers. Figure 5.4(d) presents the estimate formed using the same filter, but this time applied to the sequences of values formed by ratioing the peak pulse received power to the peak pulse transmitted power for each pulse pair. Although signal to additive noise ratios are clearly important in correcting for bias, the additive noise is generally small compared to multiplicative speckle noise. For this reason, the additive noise terms of (5.13) are ignored to simplify the subsequent analysis of potentially important systematic errors. With this modification the estimator of (5.13) becomes

$$\log_e \left[\frac{V_N(\lambda_1)}{V_N(\lambda_2)} \right]_i = \log_e \left[\frac{\bar{P}_N(\lambda_1)}{\bar{P}_N(\lambda_2)} \frac{S(\lambda_1)}{S(\lambda_2)} \right]_i \qquad (5.15)$$

where $\bar{P}_N = \bar{P}_R / P_T$.

Since $\bar{P}_N$ is the normalised power ratio uncorrupted by speckle, the lidar equation, (3.7), can be substituted into (5.15) so that

$$\log_e \left[\frac{V_N(\lambda_1)}{V_N(\lambda_2)} \right]_i = \log_e \left[\frac{D_1}{D_2} \frac{C_1}{C_2} \frac{e^{-A_1}}{e^{-A_2}} \right]_i + \log_e \left[\frac{S_1}{S_2} \right]_i \qquad (5.16)$$

where $D = \rho A_R / \pi R^2$ and $A = 2 \int_o^r \alpha r \, dr$. Further expansion yields

$$\log_e \left[\frac{V_N(\lambda_1)}{V_N(\lambda_2)}\right]_i = \log_e \left[\frac{D_1}{D_2}\right]_i + \log_e \left[\frac{C_1}{C_2}\right]_i + \log_e \left[\frac{e^{-A_1}}{e^{-A_2}}\right]_i + \log_e \left[\frac{S_1}{S_2}\right]_i \quad (5.17)$$

$$= \log_e \left[\frac{\rho_1}{\rho_2}\right]_i + \log_e \left[\frac{A_{R_1}}{A_{R_2}}\right]_i + \log_e \left[\frac{C_1}{C_2}\right]_i + [A_2 - A_1] + \log_e \left[\frac{S_1}{S_2}\right]_i \quad (5.18)$$

For a perfect lidar system, each of the log terms containing ρ, A_R, C and S would be zero leaving the differential absorption term A_2-A_1 which, upon substitution into (3.29), would yield the correct value of the instantaneous concentration at measurement time i (assuming the absorption coefficient is known exactly and no interfering species are present).

5.2.2.1 Variations in Target Backscatter

The topographic target we were constrained to using is illustrated in the photograph of Figure 5.5 The target itself is the "o" in the word "Croda" and is estimated to be between 0.5 and 1.0 metres in diameter. It is painted on a corrugated surface which is not perpendicular to the path of the laser beam. Consequently, any variation in either (a) the instantaneous area of illumination, or (b) the position of the centre of the illumination, is likely to result in variations in ρ.

Variations of type (a) will occur on a pulse to pulse basis due to changes in the refractive state of the atmosphere. unless the atmosphere can be con-

sidered to be frozen between transmission of the measurement and reference pulses, and both lasers are tuned to the same emission line. Pulse pair separation was maintained at 50 μsec (± 5 μseconds) during most of the experiments, which is a sufficiently small separation time to satisfy the frozen atmosphere requirement. However, a wavelength difference of only one line separation (the minimum achievable with a line tunable lidar system) is sufficient to decorrelate the reference and measurement channel almost completely. Typical values of the correlation coefficient estimated over 10,000 normalised power pulse pairs are approximately 0.9 where $\lambda_1 = \lambda_2$ and 0.1 where $\lambda_1 \neq \lambda_2$. Table 5.1 lists various correlation coefficients for a range of data files. Wavelength differences will be at least of the order of one CO_2 emission line separation for differential absorption measurements, and it is therefore to be expected that the speckle pattern will alter significantly between pulses in each pulse pair.

Changes in ρ due to (b), however, are most likely to occur if the beam path becomes misaligned with the target during the course of the experiment. Possible causes of misalignment arise directly from the siting of the dual hybridised laser and the transceiver optics in different rooms. Because of the long path length between the dual laser source and the transceiver (at least 4 metres) such an arrangement means the equipment will be particularly sensitive to relative motion between the two rooms. This motion may be due to:

(i) Vibration, a possible contributor here is a lift shaft which operates directly beneath the laboratories, and/or

(ii) Temperature effects both within the cavity and the optical rail system carrying the reflective optics, which would contribute a longer term component to any beam motion.

In Section 3.4 it was assumed that $\rho 1 = \rho 2 = \rho$ since a paper aperture, placed at the common focal plane of the two laser outputs and the transmitting telescope, revealed no change in the position or size of the "burn-through" hole relative to the target image over the sequences of 10,000 pulse pairs.

5.2.2.2 The Receiver Area

For a co-axial bistatic lidar system employing a single transmitting telescope, the receiver aperture, A_R, is clearly the same for both wavelengths. The $\log_e$ term in (5.18) containing A_R therefore disappears.

5.2.2.3 Efficiency of the Transceiver Optics

The output from the two hybridised pulsed laser sources must be optically combined so that both beams are co-axial at the transceiver optics. Differences in the path length thus introduced, the presence of refractive and reflective optics, and diffracting apertures will result not only in differing beam waists at all points within the system but also in beam waists which differ both at the target and also throughout the propagation path between lidar site and target.

A computer program was produced to calculate the beam diameters throughout the system starting from the grating of each hybrid laser and propagating forward to the target. Figures 5.6(a) to (d) plot these diameters in four stages; from the grating to the output coupler, the output coupler to the transmitter input lens, then to the transmitter mirror, and finally out to the target. The locations of the important

components are indicated, together with the fractional energy losses expected at each component assuming a simple circular obscuration of the beam Gaussian profile. These values are not intended as an accurate guide to energy losses but were used as an aid in identifying the principal components at fault. Figure 5.6(c), for example, reveals that the largest proportion of the losses are associated with the transmitter telescope flat as the beam is expanding on its way to the transmitter mirror. The flat also removes a smaller percentage of energy from the beam as it finally exits the transmitter on its way to the target (Figure 5.6(d)).

Although the transmitted energy was monitored as close to the exit point in the system as was practical, the primary loss inducing component was obviously sited beyond this point (see Chapter 2.0). Providing each beam remained aligned throughout each experiment, a constant correction factor could probably be applied to remove any error in the normalised power. However, the vibration and temperature effects itemised in Section 5.2.2.1 would also shift the beam profiles by different amounts relative to the obscuring components and thereby induce unknown errors in the measured value of the transmitted energy. The magnitude and significance of these errors will be analysed in Chapter 7.

5.2.3 Speckle

Historically, conventional radar analysis was first to make use of the Chi-square distribution for modelling the statistics of fluctuating radar cross-sections [11]. This distribution has the general form

$$p(x) = \frac{m}{(m-1)!\ \bar{x}} \left(\frac{mx}{\bar{x}}\right)^{m-1} \exp\left(\frac{-mx}{\bar{x}}\right), \quad x > 0 \qquad (5.19)$$

where $\bar{x}$ is the mean. According to standard statistical definitions, 2m represents the number of degrees of freedom and is always integer, but when applied to target cross-section models, m can be any real positive number. Selecting m = 1 yields the exponential or Rayleigh power distribution but, as m increases, the median of the distribution parts from the ordinate axis and tends towards the normal or Gaussian distribution. It has been found that, in radar applications, these distributions may not always fit the observed data exactly but they are usually reasonable approximations and the same has been found to be true for speckle statistics. Speckle power-frequency histograms of the Hull lidar data often revealed distributions almost identical to the Chi-square form, having values of m approximately equal to 3 or 4 (see Figure 5.2).

As indicated earlier, via expressions (5.7) and (5.8), the speckle distribution is required to have a mean of unity. This is also necessary to satisfy the requirement that the signal standard deviation vary linearly with the mean optical power received at the detector so that the variance falls to zero as the signal mean decreases to zero: i.e.,

$$\sigma^2(P_R) = \bar{P}_R^2 \, \sigma 2(S) \tag{5.20}$$

where σ^2 denotes variance. The normalised variance of the speckle noise,

$$\sigma^2(S) = \frac{\sigma^2(S)}{\bar{S}^2} = \sigma^2(S) \tag{5.21}$$

should be constant for a given receiver area and a fixed target area.

As the receiver area increases, the performance of direct detection systems improves due not only to the extra radiation imaged at the detector but also to the larger number of speckle lobes over which the optical energy is distributed. Shot to shot variations in the actual speckle count at the receiver become less significant because the addi-

tion or subtraction of a single speckle contributes relatively less to the total integrated power at the detector. This is known as "aperture averaging" since the normalised speckle variance decreases with increasing receiver area. Conversely, reductions in receiver area tend to the limit of a single speckle lobe and this corresponds to the situation for heterodyne detection where coherence requirements necessitate single mode local oscillator mixing with a single speckle [12].

If the Chi-square distribution is accepted as an adequate model, the speckle log term in (5.18) represents the difference between two log-Chi square variances, each having a mean of zero only in the limit of zero variance. Typical normalised variances, however, were of the order of 0.25 to 0.3 so bias will be present and this log ratio term will contribute an error unless the variances of both speckle sequences are identical. Differing variances could arise from disparities in the mean speckle count averaged over the active surface error of the detector.

5.2.4 Covariances

The measurement sensitivity expression, (5.1), indicates that any positive correlation between channels, for any of the parameters appearing in the lidar equation, will actually decrease the sum of the variances in each channel for each of the parameters estimated. In other words for every term, X, identified in (5.18) the error contribution in both channels will be,

$$\frac{\sigma^2(X_1)}{\bar{X}_1^{\,2}} + \frac{\sigma^2(X_2)}{\bar{X}_2^{\,2}} - \frac{2\,\mathrm{Cov}(X_1,X_2)}{\bar{X}_1,\bar{X}_2} \qquad (5.22)$$

Again, where log terms are involved, this expression is only accurate in the limit of the variance tending to zero. If, in addition to cross

channel correlation, "cross parameter" correlation also exists, then extra terms of the form + or - 2 Cov $(X_i, Y_j)/X_i Y_j$ must be included, where X and Y denote different parameters, and i,j = 1,2. These terms were included in (5.1) since the transceiver efficiencies, C_1, and C_2, are dependent on many optical components common to both channels.

The system parameters, ρ and C, are assumed to possess time dependent characteristics which have components with timescales both long and short compared to the duration of the experiments (15-20 minutes). In principle it would be desirable to have some time dependent estimate of their values, together with an indication of the variances and covariances involved. These values could then be introduced directly into (5.22) to provide a time dependent assessment of the contribution to measurement error. In practice it is difficult to provide even a simple average for these values, (formed over the duration of the experiment) since all other system parameters would have to be defined, including the 1.8 km atmospheric path and a non-ideal topographic target which is not owned by the University.

The normalised speckle variances, estimated to be of the order of 0.25 to 0.3 in the previous section, were established using a method discussed in Chapter 7. This method uses a simple, sub-optimum estimate of the time dependent means of the normalised return powers to recursively estimate the percentage normalised standard deviation. In interpreting these values directly as speckle variance, it is implicitly assumed that the contributing variances of other lidar equation parameters is negligible compared to that due to speckle.

For both channels the combined normalised speckle variance will obviously be of the order 0.5 to 0.6 but this will clearly be reduced if any correlation exists between the channels. Table 5.1 indicates

that, where identical laser lines have been selected, correlation coefficients in the range 0.8 to 0.9 are typical, but fall to values in the range 0.5 to 0.2 approximately when different lines are used. Differential absorption measurements using line tunable laser sources will therefore benefit to some extent from some correlation between the two channels. The development of suitable continuously tunable laser sources, however, which permit minimum wavelength separations between the channels will reduce the impact of speckle variance considerably in future DIAL systems.

The remaining terms in (5.18), yet to be considered, are the path integrated absorptions A_1 and A_2. In an ideal situation in which there are no interfering species and no absorption feature overlapping with the reference channel emission line, the term A_1 disappears since, $\alpha_1 = 0$. The differential absorption (A_2-A_1) of (5.18) would then simply reduce to the path integrated absorption, A_2, at the measurement wavelength λ_2. The time dependency of A_2 would probably depend not only on the origin and distribution of the gas being monitored, but also on changes induced by atmospheric turbulence and wind. In other words there will exist a power spectrum characteristic of the time dependence of A_2, and discrete samples of this path integrated absorption would form a "time series" (see next chapter) with wholly or partially random components and a set of statistical moments dependent on the dynamic state of the atmosphere.

In reality, interfering species will be present, and two of the most important contributors will be water vapour and carbon dioxide. Both have absorption features at all line tunable wavelengths, but water vapour is particularly variable both in a temporal and spatial sense, especially near coastal regions (Hull lies on the River Hull and on the Humber Estuary!). It is therefore not unreasonable to expect spatial

variations over the 1.8 km between lidar site and target, and temporal variations with time periods both long and short compared to the durations of the experiments.

Other contributions to the absorption and/or differential absorption power spectrum will arise from locally generated industrial pollution and variations in temperature and pressure.

In general, if the absorption A is composed of components, a_1, a_2, a_3, ..., a_n, where a_1 denotes absorption due to the species under investigation and, a_2 to a_n, the absorption due to interfering species, then the covariance between the path integrated absorption in each channel is given by

$$\mathrm{Cov}(A_1,A_2) = \mathrm{Cov}\left[\left(\sum_{j=1}^{N} a_1,j\right), \left(\sum_{k=1}^{N} a_2,k\right)\right] \tag{5.23}$$

Since both channels may be considered spatially and temporally coincident, full correlation would be expected between channels where $j = k$ (indicating that the same molecular species was involved) unless temperature and pressure changes effect different absorption features within the same species by different amounts. Temperature and pressure variations, however, may also result in partial correlations within and between channels whenever $j \neq k$.

In practice it would be impossible to determine the magnitude of these various covariance terms during the course of an experiment without simultaneously probing the atmospheric channel for absorption due to interfering species, and for temperature and pressure variations. This would require a multi-wavelength system.

REFERENCES

[1] BARFORD, N.C. "Experimental Measurements: Precision, Error and Truth", Addison Wesley, 1967.

[2] PETHERAM, J.C. "A comparison between continuously tunable and line tunable CO_2 laser systems for the remote sensing of atmospheric trace gases". Report prepared at Hull University (Department of Applied Physics), England, for the National Physical Laboratory, Teddington, England, 1981.

[3] Available from AFGL Hanscom Field, Bedford, MA 01730.

[4] PETHERAM, J.C. and VAN DIJK, J.W. "A high pressure laser for lidar measurements of pollutants in the atmosphere", Conference Proceedings, Eleventh International Laser Radar Conference, Wisconsin, U.S.A., June 1982.

[5] VAN DIJK, J.W., PETHERAM, J.C. and RYE, B.J. "A multi-atmosphere carbon dioxide laser for use in differential absorption lidar (DIAL)", SPIE Conference on Coherent Infrared Radar Systems and Applications II, Arlington, Vancouver, April 1983.

[6] VAN DIJK, J.W., KUSTERS, J.F., LAYFIELD, A. and RYE, B.J. "The use of TEA and multiatmosphere CO_2 lasers in active remote sensing", Proc. ESA Workshop on Space Laser Applications and Technology, Les Diablerets, ESA report SP-202, pp. 225-30.

[7] HARDESTY, R.M. "A comparison of heterodyne and direct detection CO_2 DIAL systems for ground based humidity profiling, NOAA TM ERL WPL-64, Oct 1980.

[8] RYE, B.J. "Power ratio estimation in incoherent backscatter lidar; heterodyne receiver with square law detection", J. Climate and Applied Meteorology, 22(11), Nov. 1983, pp. 1859-1913.

[9] FLUECK, J.A. and HOLLAND, B.S. "Ratio estimation and some inherent problems in their utilisation", J. Applied Meteorology, 15(6), June 1976, pp. 535-543.

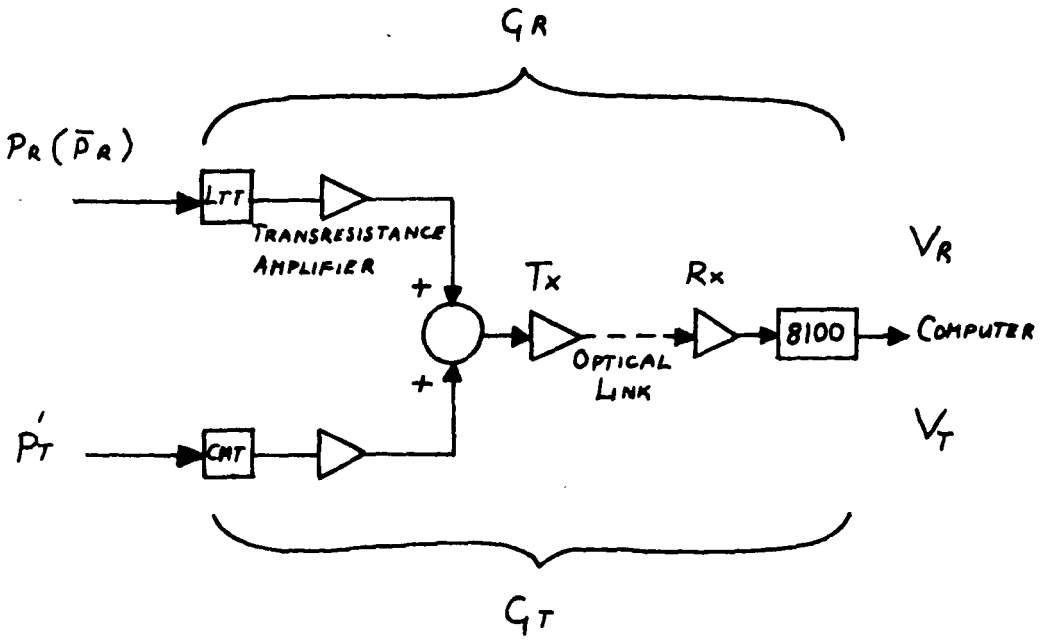
[10] KUSTERS, J.F., LAYFIELD, A.M., RYE, B.J. and VAN DIJK, J.W. "Broadband CO_2 laser and software filtering in DIAL", Proc. Workshop on DIAL Data Analysis and Control, Virginia Beach, U.S.A, 1986.

[11] SKOLNIK, M.I. "Introduction to Radar Systems", McGraw Hill, 1980.

[12] CUMMINS, H.Z. and SWINNEY, H.L. "Light Beating Spectroscopy, Chapter 3, Progress in Optics, 8, 1970.

TABLE 5.1

Filename	Gas	Wavelengths Used		Correlation Coefficients		
		Measurement	Reference	(Tx1, Tx2) (Digital Values)	(Rx1, Rx2) (Digital Values)	$\left(\frac{Rx1}{Tx1}, \frac{Rx2}{Tx2}\right)$
SIFT5.024	System Test	10 R(18)	10 R(18)	0.28	0.82	0.80
SIFT5.026	System Test	10 R(18)	10 R(18)	0.81	0.91	0.90
SIFT5.027	System Test	10 R(18)	10 R(18)	0.86	0.91	0.92
SIFT5.032	System Test	10 R(18)	10 R(22)	0.68	0.24	0.22
SIFT5.033	System Test	10 R(18)	10 R(14)	0.43	0.22	0.20
SIFT5.036	Carbon Dioxide (CO_2)	10 R(16)	10 R(8)	0.52	0.09	0.08
SIFT5.037	Ammonia (NH_3)	10 R(8)	10 R(12)	0.71	0.10	0.06
SIFT5.038	Ethylene (C_2H_4)	10 P(14)	10 P(28)	0.34	0.07	0.08
SIFT5.043	Water (H_2O)	10 R(20)	10 R(18)	0.64	0.22	0.17
SIFT5.044	Water (H_2O)	10 R(20)	10 R(18)	0.49	0.12	0.12
SIFT5.045	Water (H_2O)	10 R(20)	10 R(18)	0.56	0.12	0.09
SIFT5.046	Water (H_2O)	10 R(20)	10 R(16)	0.52	0.23	0.22
SIFT5.048	System Test	10 R(16)	10 R(18)	0.52	0.12	0.08
SIFT5.050	System Test	10 R(22)	10 R(10)	0.89	0.29	0.14



Signal Processing Components

Fig. 5.1

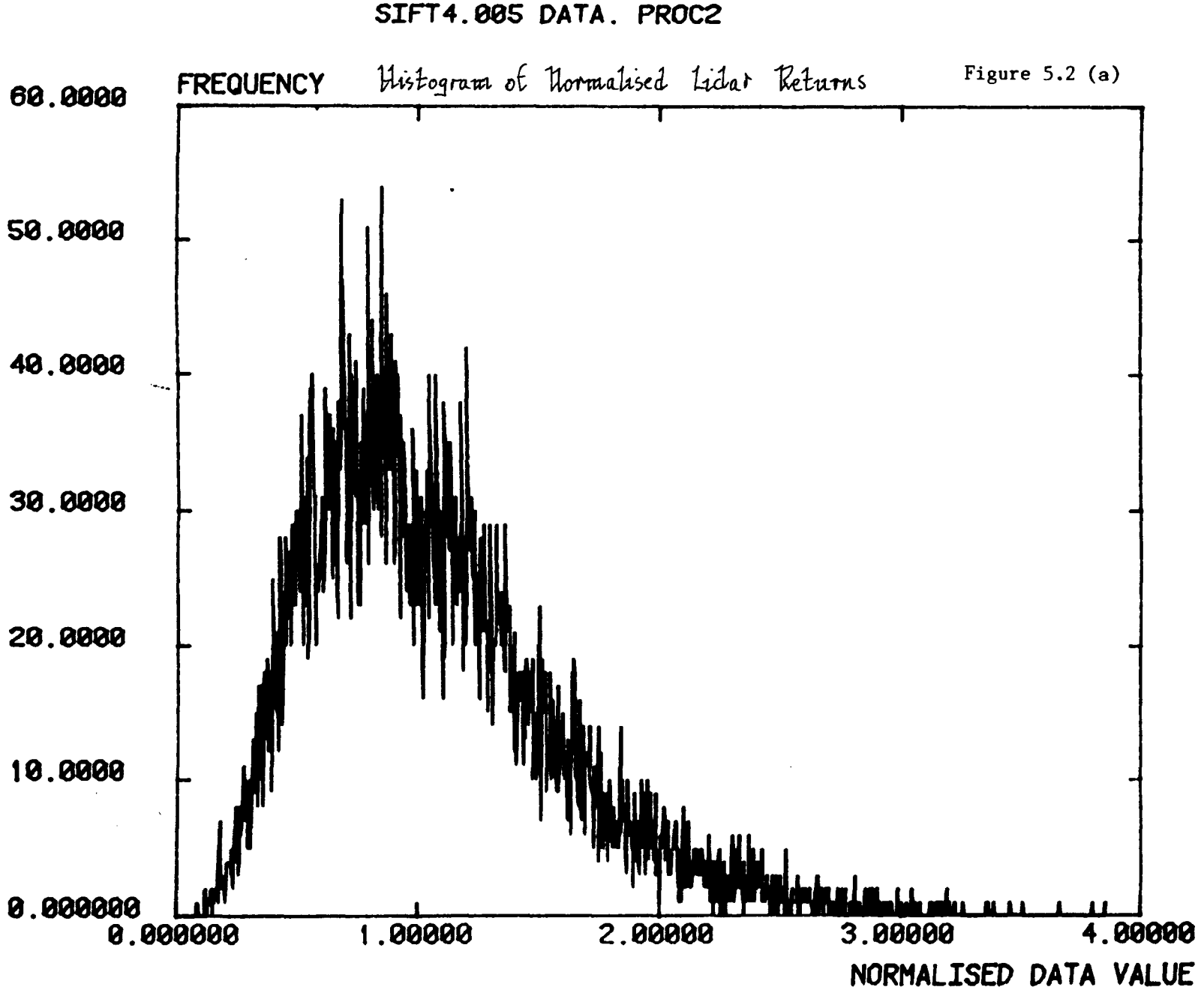

SIFT4.005 DATA. PROC2
FREQUENCY
Histogram of Normalised Lidar Returns
60.0000
50.0000
40.0000
30.0000
20.0000
10.0000
0.000000
0.000000
1.00000
2.00000
3.00000
4.00000
NORMALISED DATA VALUE

Figure 5.2 (a)

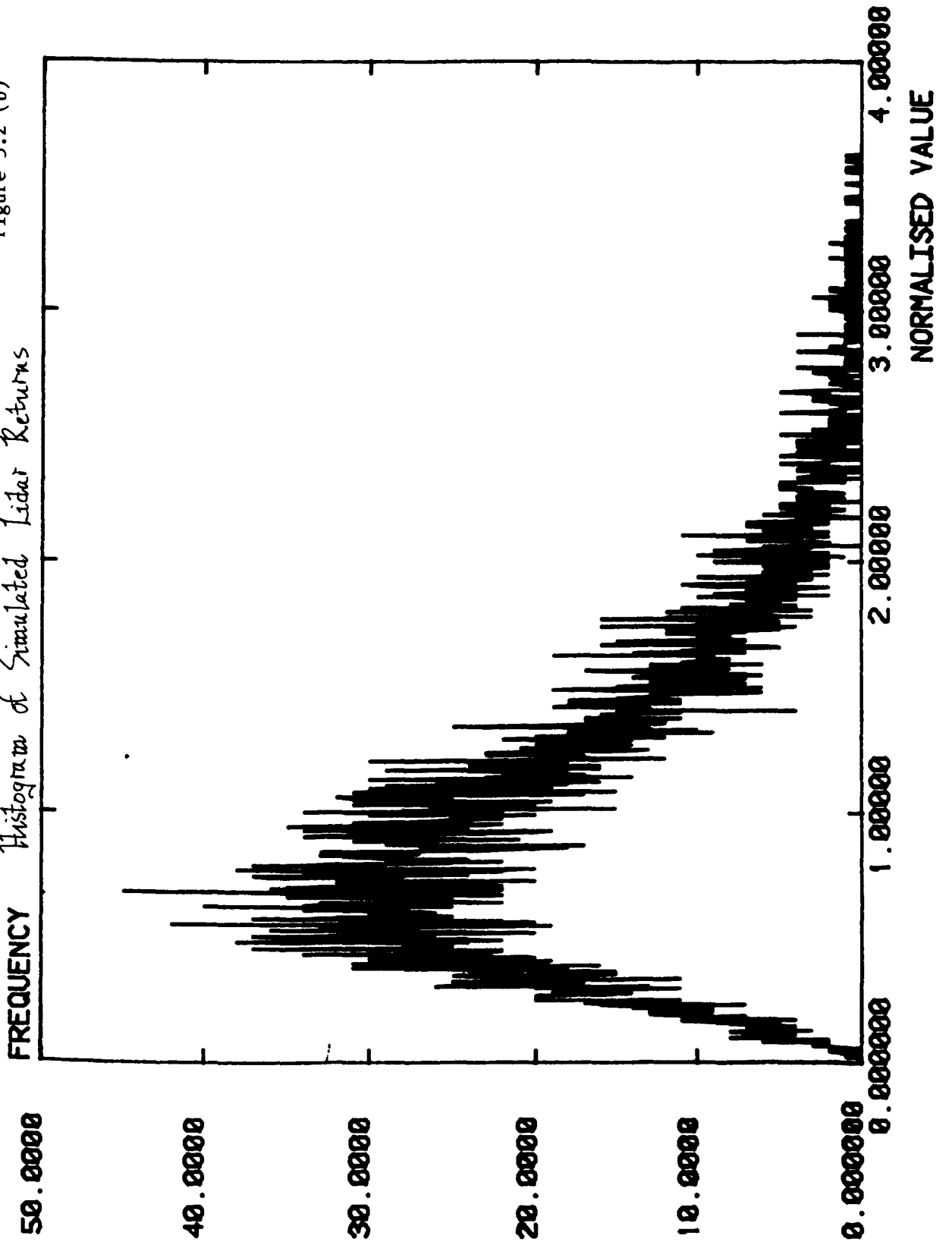

SIM (M = 3)
Histogram of Simulated Lidar Returns
FREQUENCY
50.0000
40.0000
30.0000
20.0000
10.0000
0.0000000
0.0000000
1.00000
2.00000
3.00000
4.00000
NORMALISED VALUE

Figure 5.2 (b)

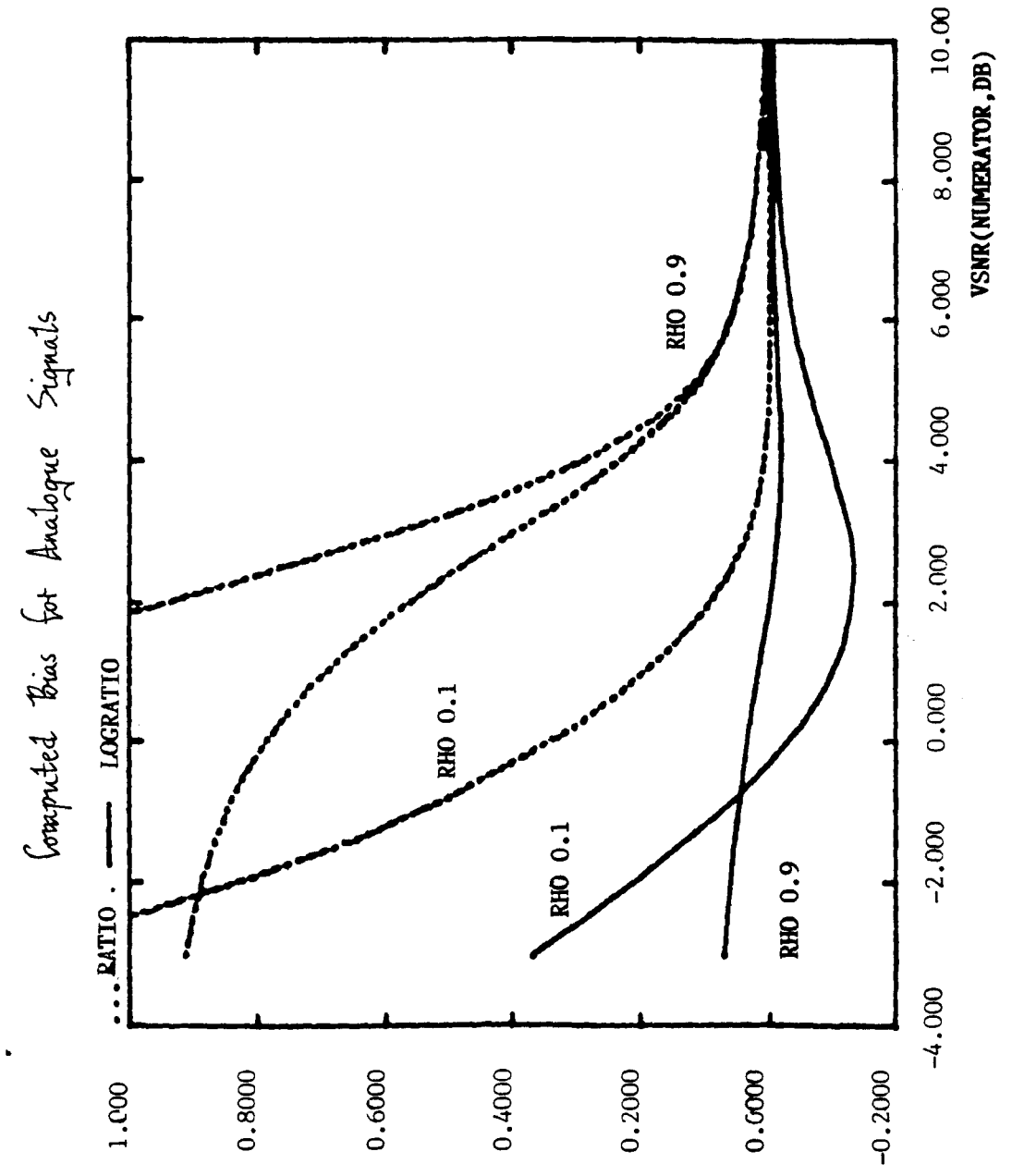
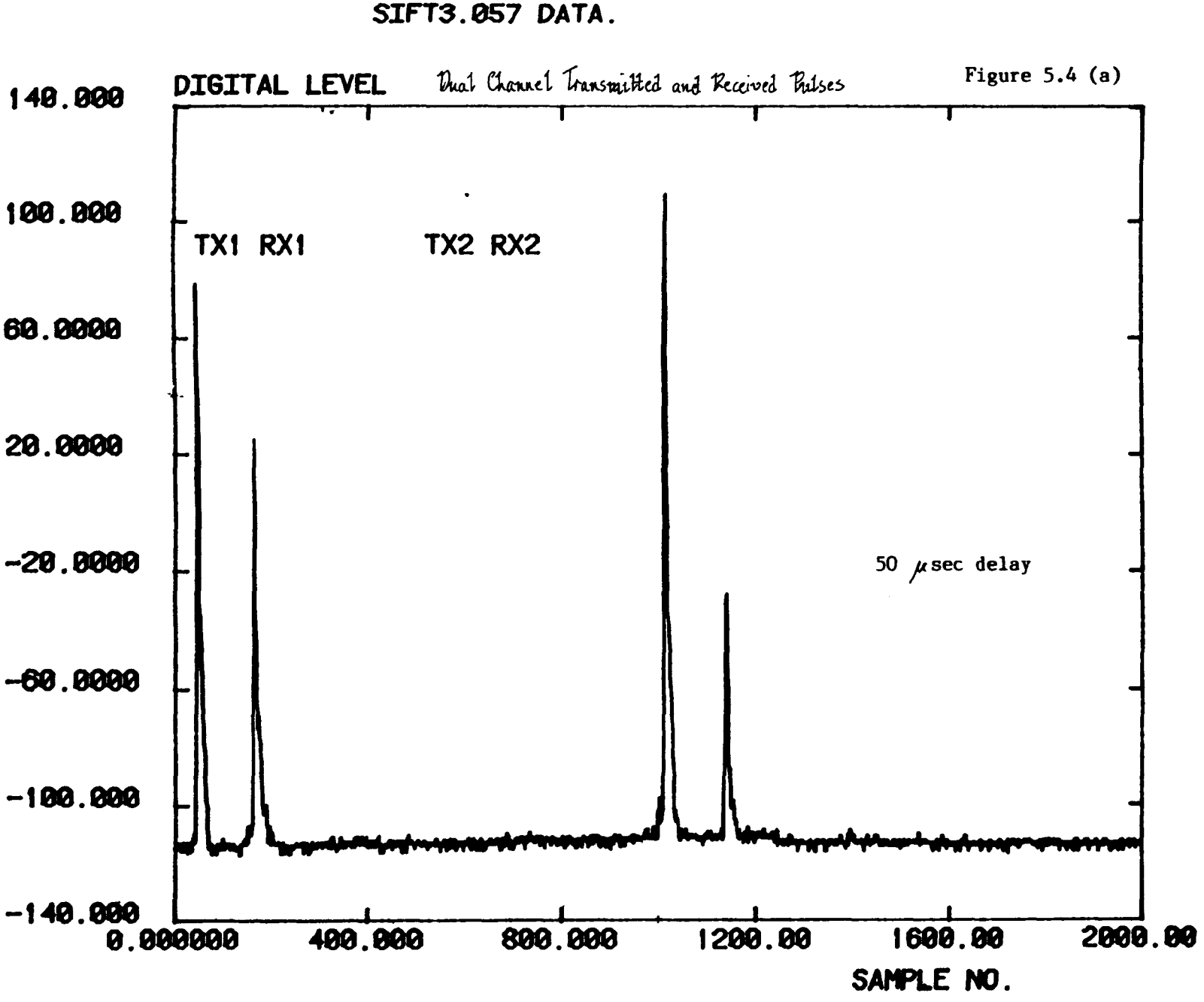

Computed Bias for Analogue Signals
....RATIO . —— LOGRATIO
RHO 0.9
RHO 0.1
RHO 0.1
RHO 0.9
1.000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-4.000
-2.000
0.000
2.000
4.000
6.000
8.000
10.00
VSNR(NUMERATOR,DB)

Fig. 5.3


SIFT3.057 DATA.
DIGITAL LEVEL
Dual Channel Transmitted and Received Pulses
Figure 5.4 (a)
140.000
100.000
60.0000
20.0000
-20.0000
-60.0000
-100.000
-140.000
TX1 RX1
TX2 RX2
50 μsec delay
0.000000
400.000
800.000
1200.00
1600.00
2000.00
SAMPLE NO.

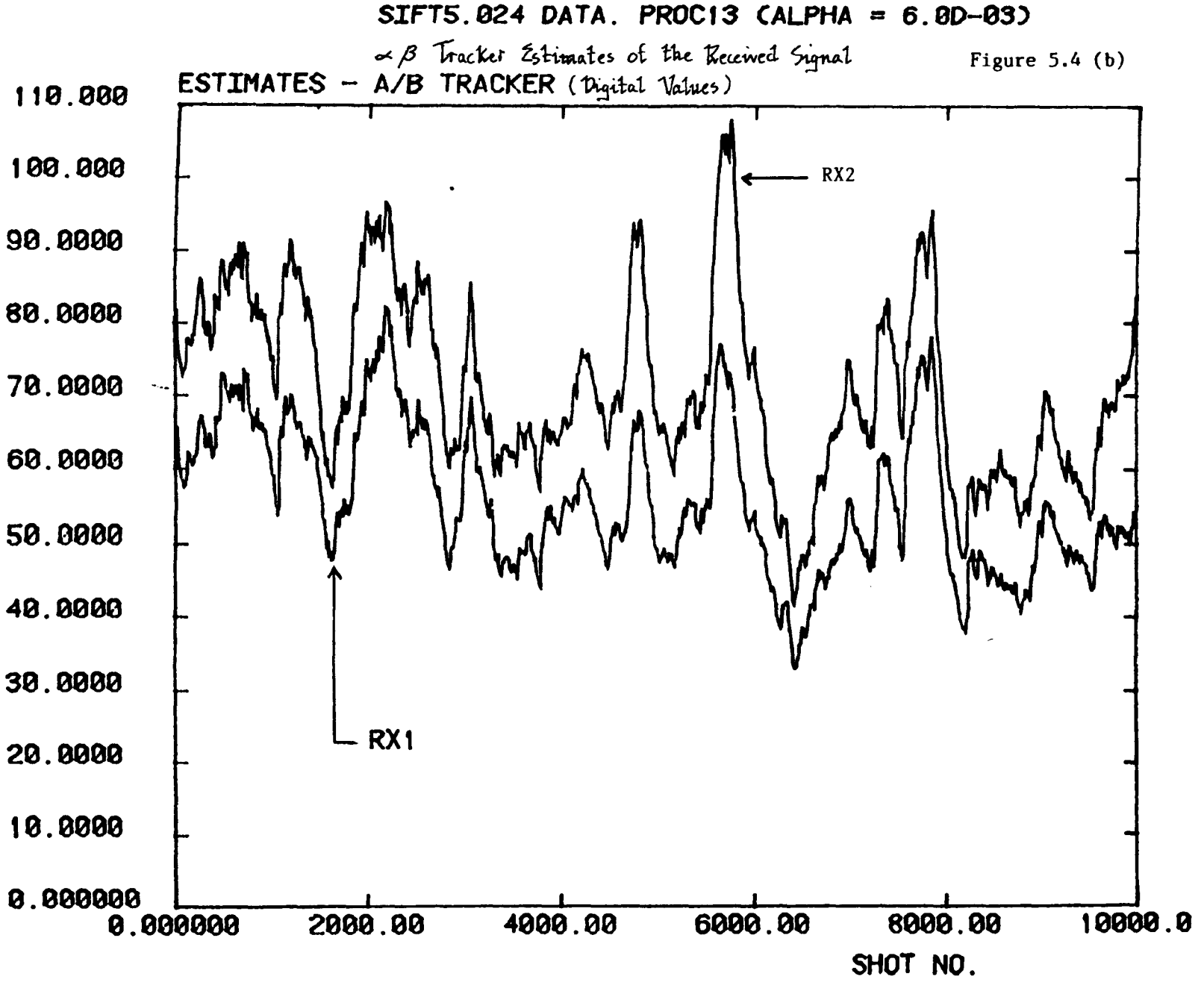

SIFT5.024 DATA. PROC13 (ALPHA = 6.0D-03)
α β Tracker Estimates of the Received Signal
ESTIMATES - A/B TRACKER (Digital Values)
110.000
100.000
90.0000
80.0000
70.0000
60.0000
50.0000
40.0000
30.0000
20.0000
10.0000
0.000000
RX2
RX1
0.000000
2000.00
4000.00
6000.00
8000.00
10000.0
SHOT NO.

Figure 5.4 (b)

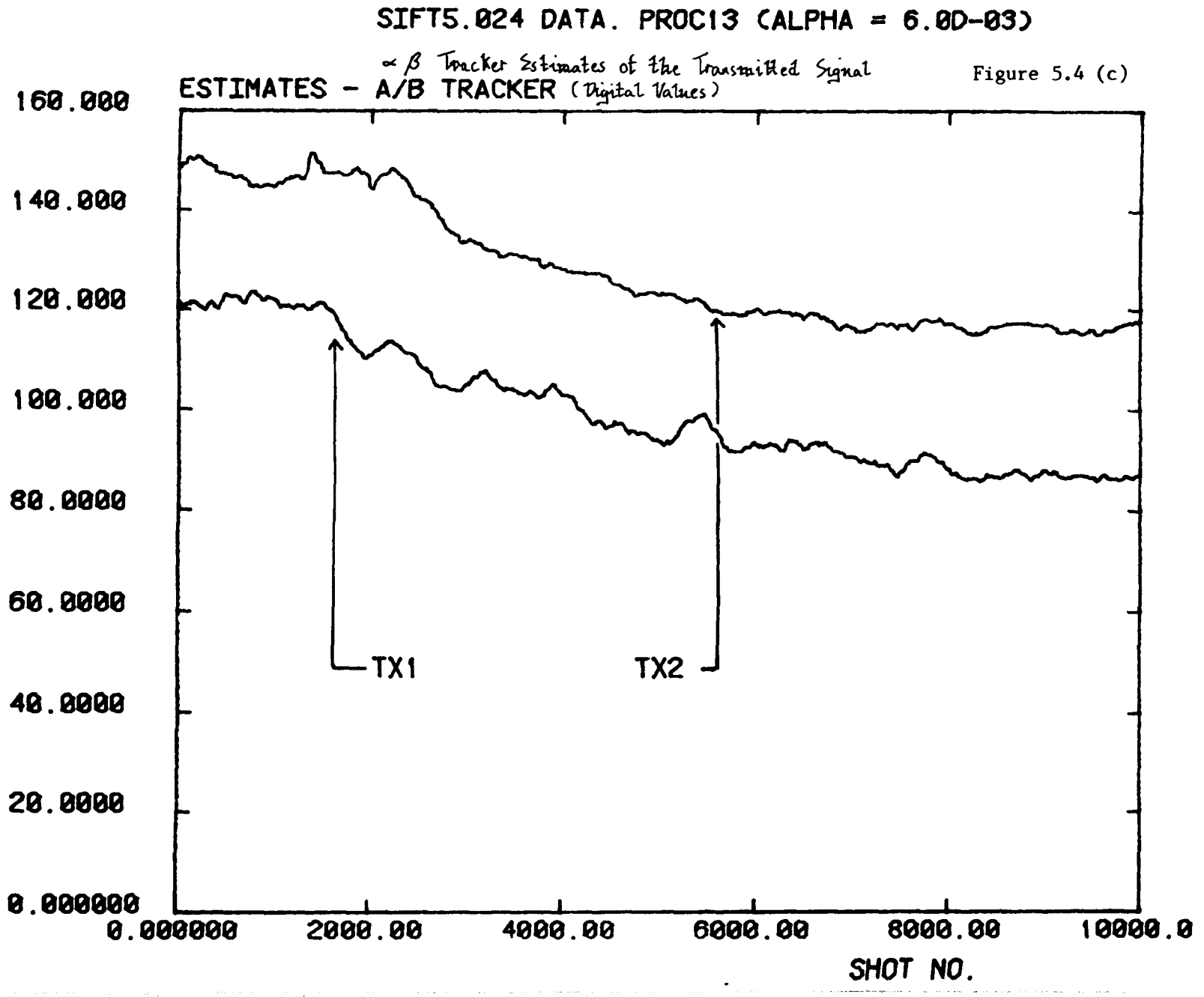

SIFT5.024 DATA. PROC13 (ALPHA = 6.0D-03)
∝β Tracker Estimates of the Transmitted Signal
ESTIMATES - A/B TRACKER (Digital Values)
160.000
140.000
120.000
100.000
80.0000
60.0000
40.0000
20.0000
0.000000
TX1
TX2
0.000000
2000.00
4000.00
6000.00
8000.00
10000.0
SHOT NO.

Figure 5.4 (c)

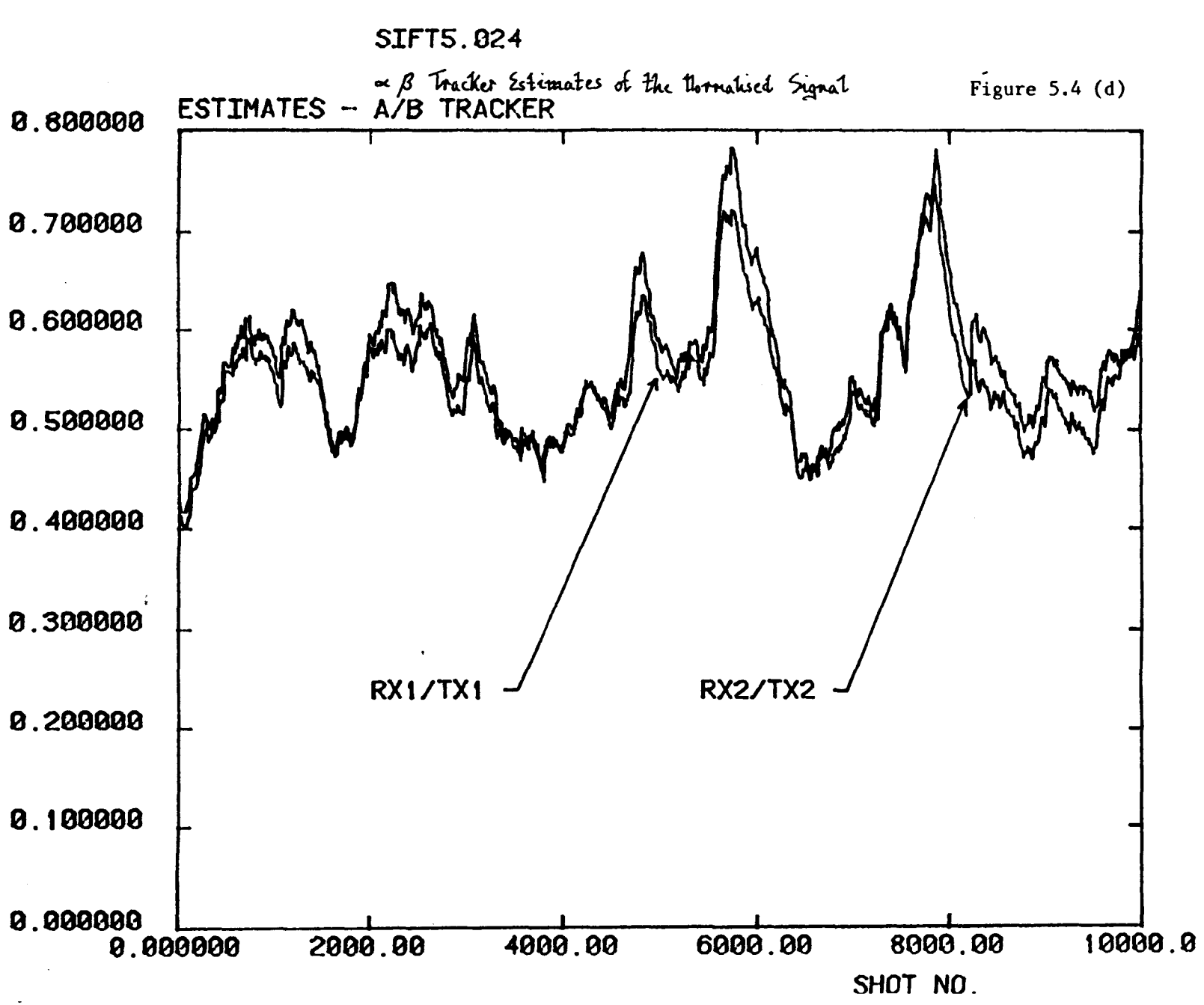

SIFT5.024
α β Tracker Estimates of the Normalised Signal
ESTIMATES - A/B TRACKER
0.800000
0.700000
0.600000
0.500000
0.400000
0.300000
0.200000
0.100000
0.000000
RX1/TX1
RX2/TX2
0.000000
2000.00
4000.00
6000.00
8000.00
10000.0
SHOT NO.

Figure 5.4 (d)

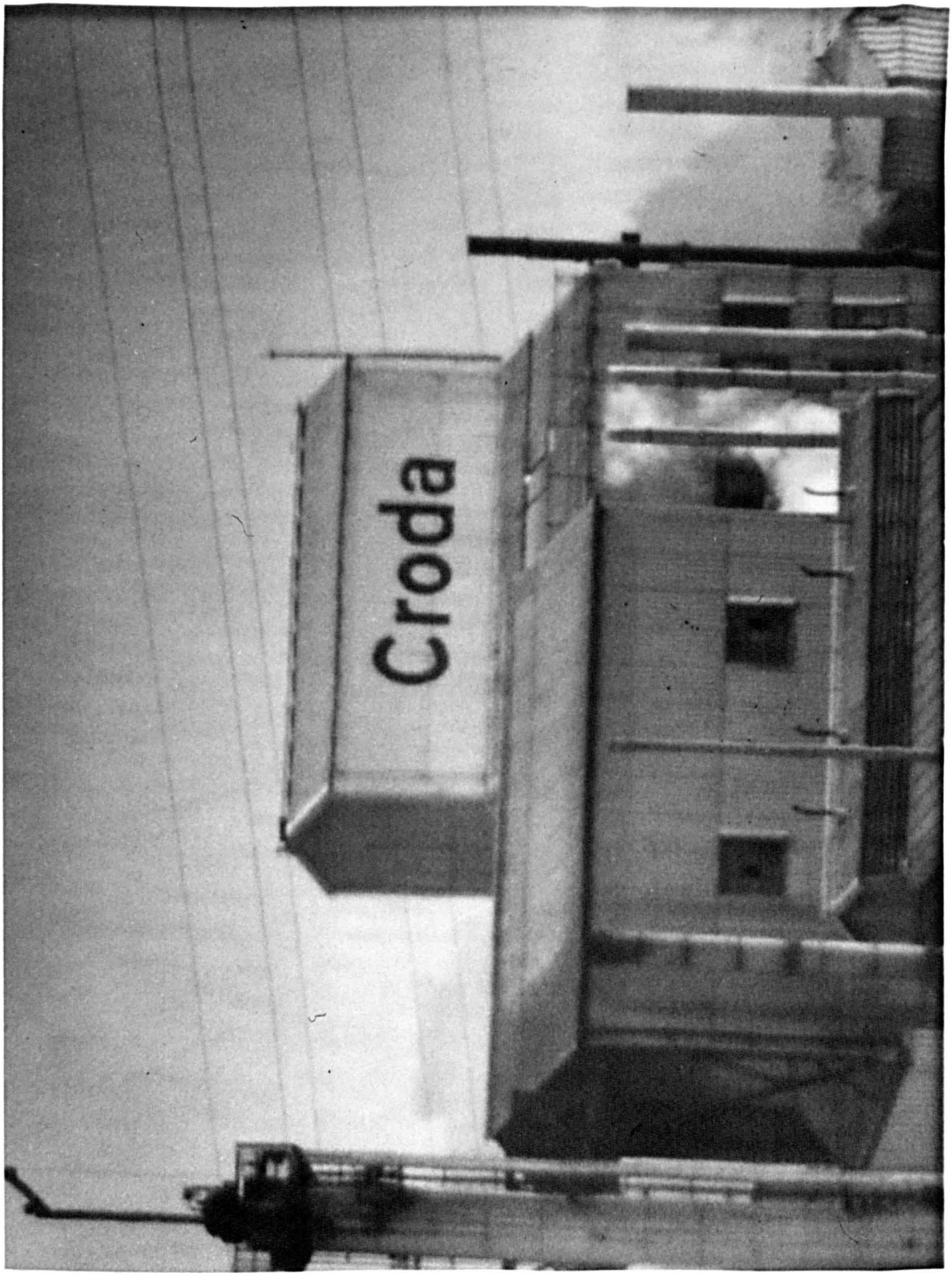


Fig. 5.5 Lidar Target

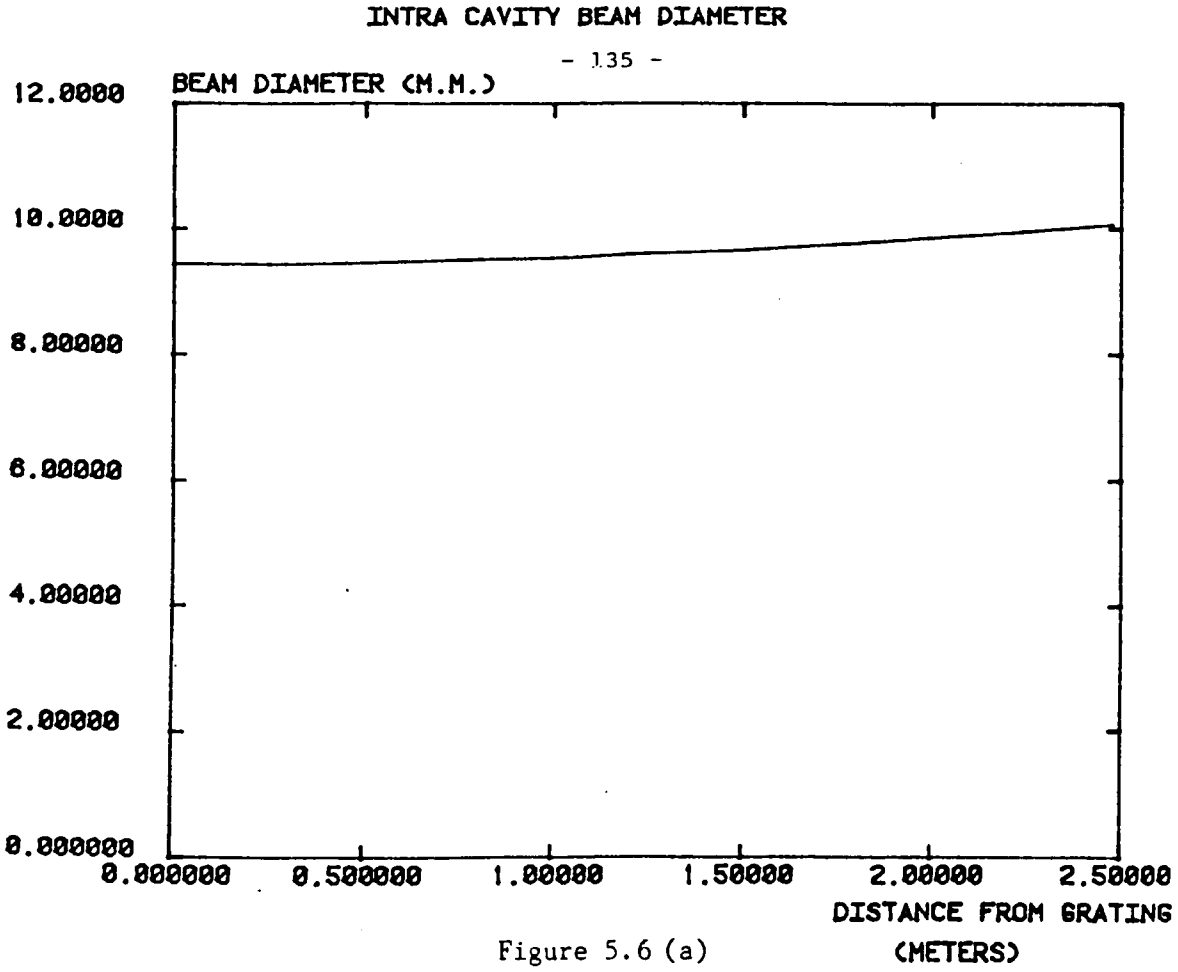


Figure 5.6 (a)

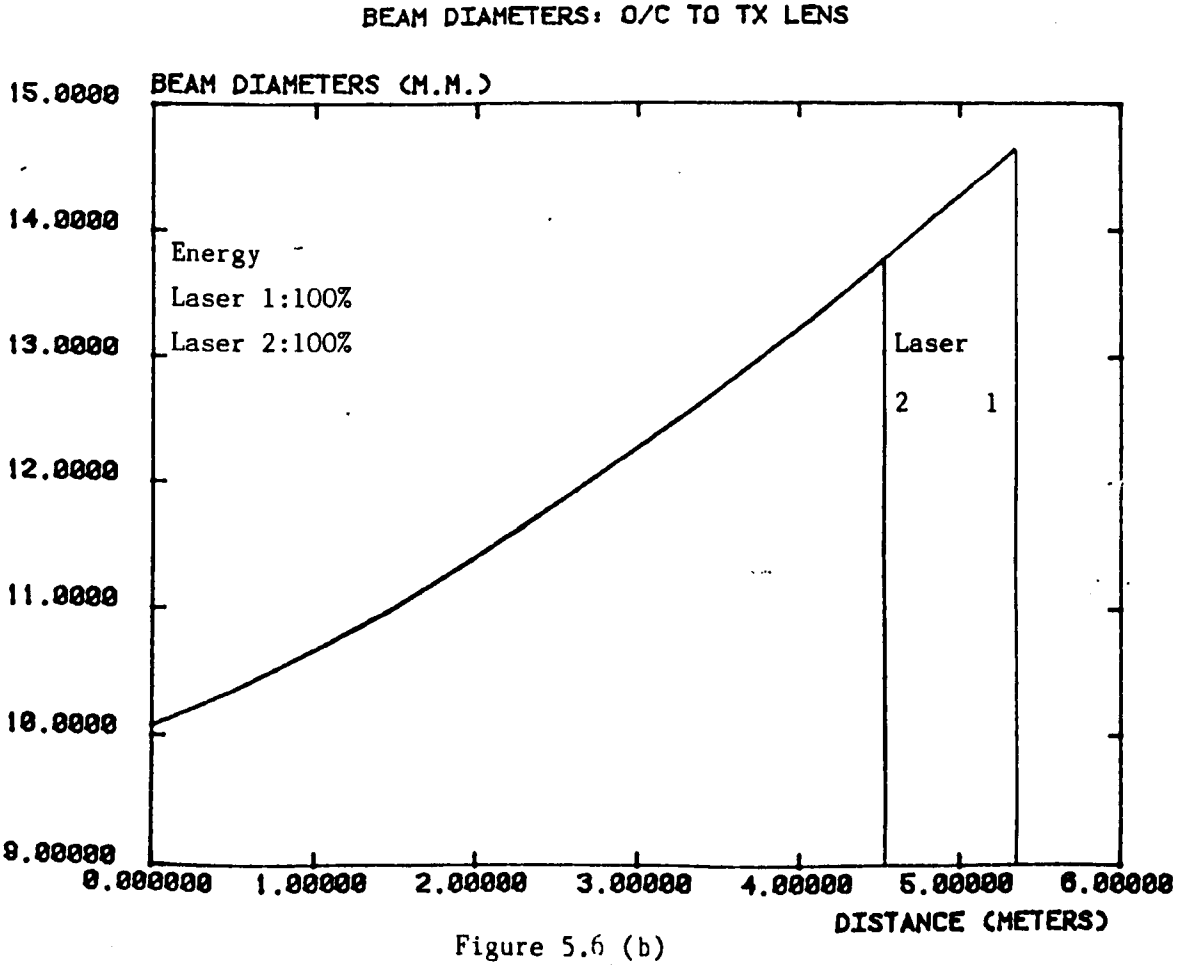


Figure 5.6 (b)

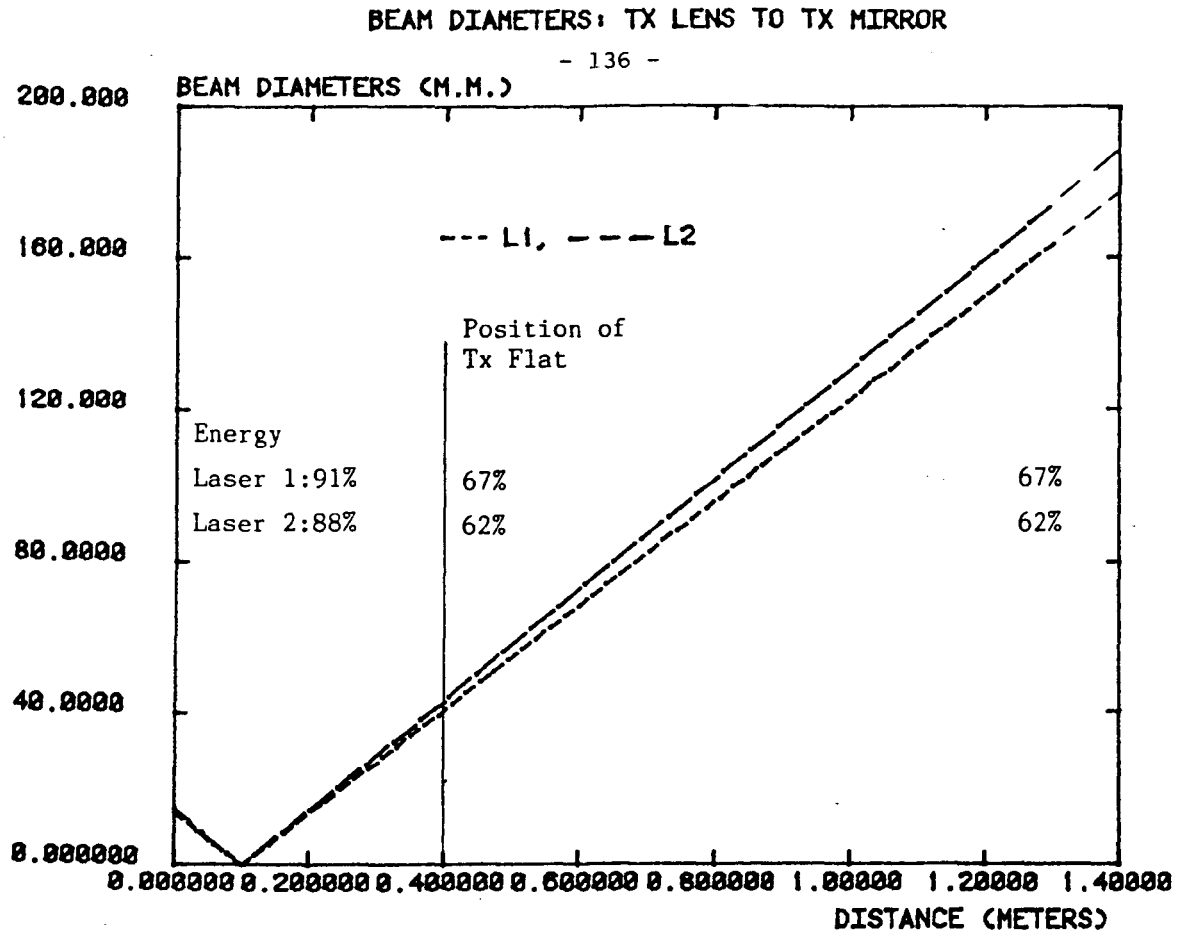


Figure 5.6 (c)

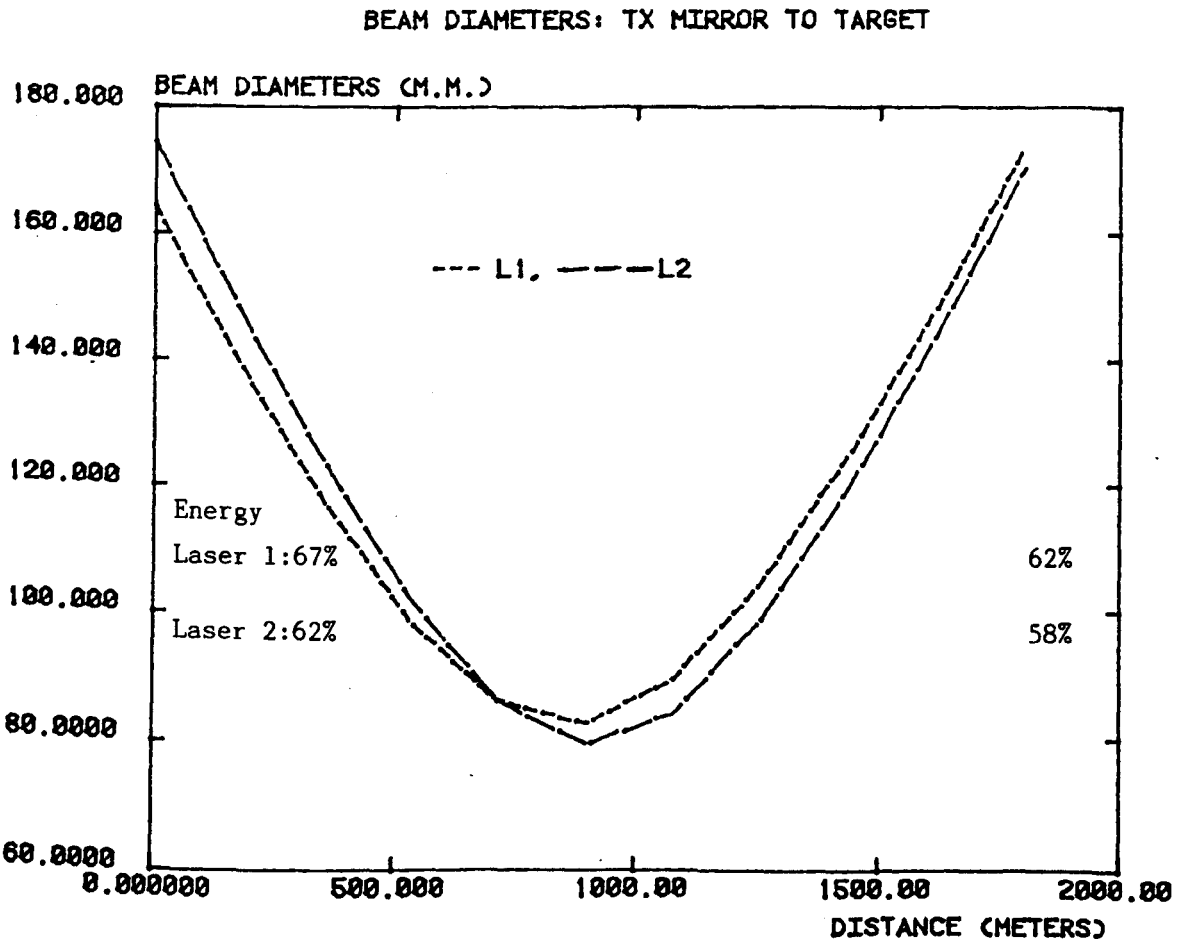


Figure 5.6 (d)

CHAPTER 6.0

ESTIMATION TECHNIQUES

Dial system accuracy is ultimately dependent upon the precision with which the log power ratio can be estimated. It is clearly important, therefore, to exploit those techniques which offer the best opportunity to improve this precision. This chapter first considers various estimation techniques which may be regarded as precursors to the methods of "optimal estimation". Selected techniques, drawn from the subject domain of optimal estimation, are then examined in an attempt to address the problem of obtaining the "best" estimate using, as criteria for determining optimality, the concepts of an unbiased, minimum variance and consistent estimator. An "unbiased" estimate is one whose expected value is the same as that of the quantity being estimated. An unbiased "minimum variance" estimate has the property that its error variance is less than or equal to that of any other unbiased estimate. A "consistent" estimate is one which converges to the true value of x as the number of measurements increases [2].

6.1 Non-Recursive Estimators

Several techniques are presented in this section which traditionally have been applied to time series type problems in general, but which have also found applications in the analysis of pulsed or sampled lidar data. Although these techniques are not in any series optimal, they are useful in providing an initial analysis of the data, including some estimate of the mean (if not the precision), and will help to reveal any trends which may exist. They are also relatively easy to implement.

6.1.1 The Batch Processor

In its most basic form, the batch processor provides a simple average. If it is assumed that a sequence of measurements, {z}, can each be expressed as a linear combination of the system parameter, x, which it is necessary to estimate, plus a random additive measurement error, then the model for the measurement process becomes

$$z(i) = x(i) + v(i) \tag{6.1}$$

where i (= 1 to k) is the sample index. The sequence {v} is assumed to have zero mean and variance, σ_v^2.

An estimate, $\hat{x}$ of x can be made using the familiar averaging technique

$$\hat{x} = \frac{1}{k} \sum_{i=1}^{k} z(i) \tag{6.2}$$

This is equivalent to summing all values of z(i) with an equal weight of 1/k. In general, however, the non-recursive filter with different weights, h(i), has the form

$$\hat{x}_k = \sum_{i=1}^{k} h(i)\, z(i) \tag{6.3}$$

If it is now assumed that x has a constant value and that some estimate of x is required from just one sample; in other words,

$$\hat{x} = z(i) = x + v(i) \tag{6.4}$$

then an error term $\tilde{x} = \hat{x} - x$ can be used to define the "mean square error", p;

$$p = E[\tilde{x}^2] = E[(\hat{x} - x)^2] = E[(x + v(i) - x)^2]$$

$$= E[v^2(i)]$$

$$= \sigma_v^2 \qquad (6.5)$$

where E denotes the expectation operator.

If, instead of a single sample, a batch of samples is now considered, then

$$p = E[\tilde{x}^2] = E[(\hat{x} - x)^2] = E\left\{\left[\frac{1}{k}\sum_{i=1}^{k}(x + v(i) - x)\right]^2\right\}$$

$$= E\left\{\left[\frac{1}{k}\left(\sum_{i=1}^{k} x + \sum_{i=1}^{k} v(i)\right) - x\right]^2\right\}$$

$$= E\left\{\left[\frac{1}{k}\sum_{i=1}^{k} v(i)\right]^2\right\}$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} E[v(i),v(j)]$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_v^2\,\delta_{ij}$$

where δ_{ij} ($\delta_{ij} = 1$ for $i = j$, $\delta_{ij} = 0$ for $i \neq j$) is the Kronecker delta symbol. This means that

$$\sum_{i=1}^{k}\sum_{j=1}^{k} \delta_{ij} = \delta_{11} + \delta_{22} + \delta_{33} + \ldots + \delta_{kk}$$

$$= k$$

and that $p = \sigma_v^2/k \qquad (6.6)$

Therefore, providing x is a constant, the mean square error decreases as k increases.

6.1.2 Segmental Averaging

This method divides the total number of samples, N, into M (= N/n) subsets of data, each containing n values. Each subset is then used to provide an estimate by forming a simple average over the n values so that

$$\hat{x}(1) = \frac{1}{n} \sum_{i=1}^{n} z_i$$

$$\hat{x}(2) = \frac{1}{n} \sum_{i=n+1}^{2n} z_i$$

$$\hat{x}(3) = \frac{1}{n} \sum_{i=2n+1}^{3n} z_i$$

$$\vdots \qquad\qquad \vdots$$

$$\hat{x}(M) = \frac{1}{n} \sum_{N-n+i}^{N} z_i \qquad (6.7)$$

For a data sequence, x, in which trends exist, this method will, in general, provide a more reasonable estimate than that formed by a simple average over all N values, because it permits the statistical properties of z to vary over the M subsets of data.

Killinger et al [1] used this as one of two estimation techniques for studying the relationship between variance and the number of samples averaged for various sequences of lidar data. In the absence of trends, this variance should decrease as 1/n. However, used as a simple estimator, and providing changes in the value of x are negligible over n data samples, the variance of each estimate is approximately

$$\sigma_I^2 = \frac{1}{n} \sum_{i=1+I}^{n} (z_i - \bar{z})^2 \qquad (6.8)$$

where I = 0, n, 2n, 3n, ..., N-n and, using (6.7), $\bar{z} = \hat{x}(I + 1)$.

6.1.3 Running Averages

The second technique used by Killinger is similar in concept but results in a larger number of estimates than segmental averaging and hence produces a more smoothly varying estimator for any non-stationary x. This method divides the total number of samples, N, into M subsets of data, each containing n values, but this time, M = N-n+1, and the estimates thus formed are

$$\hat{x}(1) = \frac{1}{n} \sum_{i=1}^{n} z_i$$

$$\hat{x}(2) = \frac{1}{n} \sum_{i=2}^{n+1} z_i$$

$$\hat{x}(3) = \frac{1}{n} \sum_{i=3}^{n+2} z_i$$

$$\vdots \qquad \vdots$$

$$\hat{x}(N-n+1) = \frac{1}{n} \sum_{i=N-n+i}^{N} z_i \qquad (6.9)$$

Unlike segmental averaging, these estimates are not independent, but each has a variance given by (6.8) providing the same restrictions that apply to segmental averaging also apply here.

6.1.4 Least Squares

The method of least squares is most frequently applied in generating a curve to fit a sequence of data points whose trend might be linear, quadratic, or of some higher order, in such a way that it minimises the error of the fit at each of the data points. By analogy with the previous definition of Section 6.1.1, the "error" is the differ-

ence between the estimate represented by the fitted polynomial and the actual (but unknown) quantity to be estimated.

Gelb [2] presents the least squares technique using vector matrix notation by first assuming that the set of n measurements, $\underline{z}$, can be expressed as a linear combination of the m elements of a constant vector $\underline{x}$ plus a random, additive measurement error, $\underline{v}$. In other words, the measurement process is modelled as

$$\underline{z} = H\underline{x} + \underline{v} \tag{6.10}$$

where $\underline{z}$ is an nxl vector, $\underline{x}$ is an mxl vector, H is an nxm matrix and $\underline{v}$ is an nxl vector.

An estimate $\hat{\underline{x}}$ of $\underline{x}$ is sought which minimises the sum of the squares of each of the deviations, $z_i - \hat{z}_i$. In vector-matrix terms this translates into the requirement that $\hat{\underline{x}}$ minimises the sum of the squares of the elements of $\underline{z} - H\underline{x}$. Since the vector inner product generates the sum of the squares of the vector, a scalar "cost function", J, may be defined of the form

$$J = (\underline{z} - H\hat{\underline{x}})^T(\underline{z} - H\hat{\underline{x}}) \tag{6.11}$$

Partial differentation of this scalar with respect to the vector $\underline{x}$ will yield the minimum when

$$\frac{\partial J}{\partial \hat{\underline{x}}} = \underline{0} \text{ for } \left|\frac{\partial^2 J}{\partial \hat{\underline{x}}^2}\right| \geq 0 \tag{6.12}$$

Performing this differentiation and setting the result equal to zero yields

$$H^T H\hat{\underline{x}} = H^T\underline{z} \tag{6.13}$$

so that if H^TH possesses an inverse, the least squares estimate becomes

$$\hat{\underline{x}} = (H^TH)^{-1} H^T\underline{z} \tag{6.14}$$

Here, again, a batch processing scheme is required since all the available measurements are utilised together at one time. Furthermore, this technique does not supply an optimal estimator in terms of the criteria defined at the beginning of this chapter, since it is not necessarily consistent, nor of minimum variance, unless the polynomial order can be selected to be as close as possible to that of the underlying system trend.

6.1.5 Maximum Likelihood and Bayesian Estimation

Least squares estimates of the form derived above are obtained using deterministic arguments only, and make no assumptions about the statistical properties of $\underline{z}$ (via e or $\underline{v}$) and $\underline{x}$. An alternative approach, therefore, may be derived from the "maximum likelihood" philosophy which defines $\hat{\underline{x}}$ as the value which maximises the probability of the measurements, $\underline{z}$, having actually occurred. A statistical model is assumed to exist for $\underline{v}$, but not for $\underline{x}$. Using the measurement model defined in the previous section, a conditional probability density function, $p(\underline{z}|\underline{x})$, is obtained which relates the probability of obtaining a particular value, $\underline{z}$ to a given value of $\underline{x}$. This is equivalent to the probability density function for $\underline{v}$ centred on $H\underline{x}$ and has the form [2],

$$p(\underline{z}|\underline{x}) = \frac{1}{(2\pi)^{n/2}|R|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}(\underline{z} - H\underline{x})^T R^{-1}(\underline{z} - H\underline{x})] \tag{6.15}$$

where $\underline{v}$ is assumed to be a zero mean gaussian random variable, having a covariance matrix, R, defined as

$$R = E[\underline{v}\ \underline{v}^T] \tag{6.16}$$

Since the requirement is to maximise $p(\underline{z}|\underline{x})$, it is necessary to minimise the exponent in square brackets. This exponent is almost identical to the least squares cost function, except for the inverse of the covariance matrix, R, which has the effect of weighting the sum of the squares of the deviations instead of treating them all as of equal significance. In this case, the least squares result given above is modified and becomes

$$\hat{\underline{x}} = (H^T R^{-1} H)^{-1} H^T R^{-1} \underline{z} \tag{6.17}$$

A further alternative exists in the "Bayesian" estimation method where statistical models are assumed for both $\underline{x}$ and $\underline{z}$. In contrast to the previous method, an "a posteriori" conditional density function of the form $p(x|z)$ is sought which is related to the previous conditional probability function via Bayes' theorem:

$$p(\underline{x}|\underline{z}) = \frac{p(\underline{z}|\underline{x})\ p(\underline{x})}{p(\underline{z})} \tag{6.18}$$

where $p(\underline{x})$ is the "a priori" probability density function of $\underline{x}$ and $p(\underline{z})$ is the probability density function of the measurements. Solutions for the estimate, $\hat{\underline{x}}$, can be obtained from $p(\underline{x}|\underline{z})$ but their exact form depends on the criteria of optimality. If the requirement is to maximise the probability that $\hat{\underline{x}} = \underline{x}$ then a solution is obtained by equating $\hat{\underline{x}}$ to the mode of the distribution. For this particular example, if $p(\underline{x})$ is assumed to be uniform (i.e. no mode at a unique value of x), then this estimate is equivalent to the maximum likelihood estimate [2].

Gelb presents the generalised minimum variance Bayes' estimate as

$$\hat{\underline{x}} = (P^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} \underline{z} \tag{6.19}$$

where gaussian distributions are assumed for both $\underline{x}$ and $\underline{v}$, and P_o is the "a priori" covariance matrix of $\underline{x}$ defined as

$$P_o = E[\underline{x}\,\underline{x}^T] \tag{6.20}$$

In comparing these estimation methods, he notes that if there is little or no "a priori" information, P_o^{-1} becomes very small and the above estimate reverts to the maximum likelihood result. Similarly, if all measurement errors are uncorrelated, so that R becomes a diagonal matrix, and if all errors have equal variance so that $R = \sigma^2 I$, where σ^2 denotes the variance of $\underline{v}$ and I is the identity matrix, then the maximum likelihood result reduces to the least squares estimate. He thereby arrives at the important conclusion that the methods of least squares, maximum likelihood and Bayesian estimation produce identical results if gaussian random variables are assumed throughout and the other assumptions are the same in each case.

Before moving on to consider recursive filters in general, the important topic of Kalman filtering in particular, and relationships that exist between recursive and non-recursive implementations, the optimum non-recursive estimator, the Wiener filter, must be considered. It is worth noting here that Kalman's important contribution [7], which makes use of both state space representations and probability theory to solve the Wiener problem for gauss-markov sequences, also reduces to the estimators given above.

6.1.6 Optimum Non-Recursive Estimator (The Wiener Filter)

By way of introduction, this section returns to the linear batch processor, first identified in Section 6.1.1, having an estimator of the general form

$$\hat{x}_k = \sum_{i=1}^{k} h(i)\, z(i) \tag{6.3}$$

In that earlier section all the coefficients, h(i), had equal weights and the estimator produced a simple average, but this section turns to the problem of selecting the coefficients in such a way that the mean square error

$$p = E[e^2] = E[(x - \hat{x})^2] \tag{6.21}$$

is minimised. Here, again, x is used to denote the desired signal and $\hat{x}$ its estimate. Following Bozic [5], substituting (6.3) into (6.21) gives

$$p = E\left[\left(x - \sum_{i=1}^{k} h(i)\, z(i)\right)^2\right] \tag{6.22}$$

The minimum mean square error is obtained by differentiating (6.22) with respect to each of the coefficients and setting the result equal to zero.

$$\frac{\partial p}{\partial h(j)} = -2E\left[x - \sum_{i=1}^{k} h(i)\, z(i)\right] z(j) = 0 \tag{6.23}$$

or

$$\sum_{i=1}^{k} h(i)\, E[z(i)\, z(j)] = E[xz(j)] \tag{6.24}$$

where j = 1, 2, 3, ..., k. By defining

$$p_z(i,j) = E[z(i)\, z(j)] \tag{6.25}$$

as the autocorrelation function of z(i) and

$$p_{xz}(j) = E[xz(j)] \tag{6.26}$$

as the cross correlation function of the random variables x and z(i), equation (6.24) can be re-written as

$$\sum_{i=1}^{k} h(i)\, p_z(i,j) = p_{xz}(j) \quad \text{for } j = 1, 2, \ldots, k \tag{6.27}$$

This expression can be expanded over both i = 1, 2, ..., k and j = 1, 2, ..., k to yield a set of simultaneous equations, each with k terms. A more compact way of presenting these equations is to adopt matrix notation whereby (6.27) becomes

$$P_z \underline{h} = p_{xz} \tag{6.28}$$

where P_z is the kxk correlation matrix, and $\underline{h}$ and p_{xz} are kxl column vectors. Solving (6.28) for $\underline{h}$ yields

$$\underline{h} = P_z^{-1} p_{xz} \tag{6.29}$$

and (6.3) can be written

$$\hat{x} = \underline{h}^T \underline{z} \tag{6.30}$$

where $\underline{h}$ and $\underline{z}$ are kxl column vectors, and $\underline{h}^T$ (the transpose of h) is a row vector.

Substituting (6.29) into (6.30) provides the estimate,

$$\hat{x} = p_{xz}^T P_z^{-1} \underline{z} \tag{6.31}$$

and the least mean square error,

$$p = E[x^2] - \underline{p}_{xz}^T P_z^T p_{xz} \tag{6.32}$$

It is noteworthy that, in common with the estimators discussed in previous sections, the dimensionality of any algorithm used to implement the filter grows with the number of samples to be processed and must include the inversion of a kxk matrix. Furthermore, the above expressions represent only the scalar case in which a single signal is to be

estimated. A more general result would consider a vector signal (multi-dimensional) and would be considerably more complex for a large number of samples.

However, since no assumptions are made regarding how the signal, x, and noise, v, are combined (for example, they need not be additive as in (6.1)), the above result is applicable to a wider range of problems than other filters discussed elsewhere in this chapter, and represents the best non-recursive linear filter operation that may be carried out on the samples in order to estimate x.

6.2 Recursive Estimators

Gelb defines a recursive estimator as one in which there is no need to store past measurements for the purpose of computing present estimates. Instead, measurements are utilised sequentially, as they become available. Recursive estimators in terms of this definition are therefore normally far preferable to non-recursive estimators of identical performance because they are so much more economical in terms of computational storage.

Kalman and others [6-10] during the 1960s were responsible originally for advancing optimal recursive estimator techniques using "state space", time domain formulations. This approach, now generally referred to as Kalman filtering, has become the primary technique for analysing and optimising data mixing in modern multi-sensor systems, since it is ideally suited to digital computer implementation.

Gelb identifies three basic types of estimation problem, depicted in Figure 6.1. According to his definition, the term "filtering" applies when the time at which an estimate is desired coincides with the last available measurement point, (a). If the desired estimate falls within

the span of available measurement data, the estimation problem is then referred to as "smoothing", (b), but if the time of interest occurs after the last available measurement then the estimated quantity becomes a prediction, (c).

It is useful to append to the above definitions a distinction between "recursive" and "iterative" data processing introduced by Young [11]. Figure 6.2 depicts recursive processing in which an estimate is obtained by working serially through the data, one sample at a time. Iterative data processing, on the other hand, refers to the sequential processing of a complete set of data in which, at each step, the data base remains the same at N samples and only some estimated variable is modified. Simple "en bloc" processing then becomes a special case in which a single iteration is considered.

6.2.1 First Order Recursive Estimators

Following the example provided by Gelb [2], a scalar non-random constant, x, which is unknown and has no defined statistical properties, and which is also corrupted by noise measurements of the form given by (6.1) may be estimated by averaging these measurements using (6.2). Reproducing these expressions,

$$z(i) = x + v(i) \tag{6.1}$$

$$\hat{x}_k = \frac{1}{k} \sum_{i=1}^{k} z(i) \tag{6.2}$$

it is noted here that they define an estimator which is both unbiased and of minimum variance.

When an additional measurement becomes available, the new estimate becomes

$$\hat{x}_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} z_i \tag{6.33}$$

and, after a simple rearrangement,

$$\hat{x}_{k+1} = \frac{k}{k+1} \left(\frac{1}{k} \sum_{i=1}^{k} z_i \right) + \frac{1}{k+1} z_{k+1} \tag{6.34}$$

$$= \frac{k}{k+1} \hat{x}_k + \frac{1}{k+1} z_{k+1} \tag{6.35}$$

$$= \left[\frac{\hat{x}_k (k+1) - \hat{x}_k}{(k+1)} \right] + \frac{1}{k+1} z_{k+1} \tag{6.36}$$

$$= \left[\hat{x}_k - \frac{\hat{x}_k}{k+1} \right] + \frac{1}{k+1} z_{k+1} \tag{6.37}$$

$$= \hat{x}_k + \frac{1}{k+1} (z_{k+1} - \hat{x}_k) \tag{6.38}$$

In this form the need to store past measurement is eliminated since all previous information is embodied in the prior estimate. Each new estimate is given by the prior estimate plus an appropriately weighted difference between the new measurement and its expected value (the prior estimate). The quantity in parentheses is generally referred to as the "residual" and usually given the symbol, ν.

Expression (6.38) is a simple averaging filter and, in principle, it is identical in function to the averaging batch processor of Section 6.1.1. In practice, however, (6.38) is far more economical in terms of variable storage when implemented as a computer algorithm. The variance reduction equation, (6.6), is clearly still valid here for constant x, but a similar analysis to the above can again be applied to the problem of estimating the variance; a recursive algorithm of the form

$$\hat{\sigma}^2_{k+1} = \hat{\sigma}^2_k - \frac{1}{k+1} \left[\hat{\sigma}^2_k - e^2_k \right] \tag{6.39}$$

is obtained where $e_k = z_k - \hat{x}_k$.

If the weighting factor, 1/(k+1), is replaced by b, then (6.38) becomes,

$$\hat{x}_{k+1} = \hat{x}_k + b(z_{k+1} - \hat{x}_k) \tag{6.40}$$

and further rearrangement will yield,

$$\hat{x}_{k+1} = (1-b)\hat{x}_k + b\, z_{k+1} \tag{6.41}$$

Both are general forms of the first order recursive filter in which b need not be a function of k, but is restricted to values in the range $0 < b < 1$. Schematically, the filter algorithms, (6.40) and (6.41) may be represented as shown in Figure 6.3 where D denotes a single delay in the recursive procedure.

Zrnic has demonstrated the application of first order recursive filters in estimating the mean power from smoothed square law detector outputs of stationary signals [3]. His signals are generated by radar echoes from solid objects or by the scatter cross-sections of randomly distributed point scatters. Their statistical characteristics are very similar to those generated by direct and heterodyne detection lidar. He states, however, that if an arbitrary value of b is selected and the square law detector samples, which are unbiased, are block averaged over N values, then the same variance reduction is obtained as for the recursive estimator above, providing

$$k = N = (2-b)/b \tag{6.42}$$

If the number of samples available exceeds N (i.e. $k > N$), then presumably the recursive estimator behaves as a "moving exponential window" (discussed later).

A discrepancy between the results obtained by Gelb and Zrnic becomes apparent if (6.42) is re-written in terms of k so that

$$b = \frac{2}{k+1} \tag{6.43}$$

The general case of the first order recursive filter, (6.40), should become identical to the specific case of the simple recursive averager (6.38) or block averager, if k = N. The values of b, however, obtained in (6.40) and (6.42) differ by a factor of 2.

Furthermore, Zrnic also states that the first order recursive filter, applied to square law detector outputs, produces an estimate which is biased by an amount $(1-b)^{k+1}$. Again, for the case where k = N, this bias would become

$$\left[1 - \frac{2}{k+1}\right]^{k+1} = \left[\frac{k-1}{k+1}\right]^{k+1} \tag{6.44}$$

if (6.42) and (6.43) are correct, which tends to a value of 0.1353 in the limit of large k. Bias will still be present even if the alternative value of b is used, and if true would seem to imply that the averaging filter of (6.38) is biased! There are, however, differences in the assumptions applied to the statistics of the filter input sequences, z. Gelb assumes a constant mean, x, corrupted by an additive, zero mean white noise sequence, v. Zrnic, on the other hand, assumes that z has an exponential power distribution of the form

$$P(z) = \exp(-z/P_o)/P_o$$

where P_o is the constant true mean power at the input of the square law detector.

Both Gelb's and Zrnic's interpretation of the first order filter

are intended for use with constant mean data. First order filters, in general, are biased for transients; bias, in fact, can only be avoided by resorting to a filter of an order equivalent to the highest differential existing within the transient. For example, to avoid the bias introduced into the estimate of a first order filter by a linear ramp transient, it is necessary to track the slope and this necessitates the application of a second order filter. To avoid bias due to quadratic transients, a third order filter is necessary, and so on.

Before examining a particular second order filter in the next section, a brief return is made to a particular nomenclature used to describe certain first order recursive estimators - the "exponential weighting, smoothing, or windowing" filter. In common with the filters discussed above, the principle is to produce an estimate based on a weighted sum of all past observations, so that

$$\hat{x}_k = c_o z_k + c_1 z_{k-1} + c_2 c_{k-2} + \ldots + c_{k-1} z_1 \tag{6.45}$$

where (c_i) are the weights. More weight is given to the most recent observations and less to observations further in the past; in order that these weights sum to zero, a geometric series is selected of the form

$$c_i = \alpha(1-\alpha)^i, \quad i = 0, 1, 2, \ldots, k-1 \tag{6.46}$$

$$0 < \alpha < 1$$

Equation (6.45) then becomes,

$$\begin{aligned} \hat{x}_k &= \alpha z_k + \alpha(1-\alpha) z_{k-1} + \alpha(1-\alpha)^2 z_{k-2} + \ldots + \alpha(1-\alpha)^{k-1} z_1 \\ &= \alpha z_k + (1-\alpha)[\alpha z_{k-1} + \alpha(1-\alpha) z_{k-2} + \ldots + \alpha(1-\alpha)^{k-2} z_1] \\ &= \alpha z_k + (1-\alpha)\hat{x}_{k-1} \end{aligned} \tag{6.47}$$

which, clearly, is identical to (6.41).

An alternative form sets

$$c_i = (1-\alpha)^i, \quad i = 0, 1, 2, \ldots, k-1 \qquad (6.48)$$

$$0 < \alpha < 1$$

to provide an estimate

$$\hat{x}_k = z_k + (1-\alpha)\hat{x}_{k-1} \qquad (6.49)$$

derived in exactly the same way as (6.47). Both weighting sequences, (6.46) and (6.49), exhibit a characteristic exponential decay as i increases from zero.

6.2.2 A Second Order Recursive Estimator: The αβ Tracker

Historically, the αβ tracker has been applied most frequently in ranging type applications where it is necessary to have an estimate not only of the range and radial velocity of some object/target, but also a prediction of where the object might be at some future time beyond the current measurement. The αβ tracker is governed by the set of equations,

$$\hat{x}_{k+1} = x^p_{k+1} + \alpha(z_{k+1} - x^p_{k+1}) \qquad (6.50)$$

$$\hat{\dot{x}}_{k+1} = \hat{\dot{x}}_k + \frac{\beta}{T}(z_{k+1} - x^p_{k+1}) \qquad (6.51)$$

$$x^p_{k+1} = \hat{x}_k + T\,\hat{\dot{x}}_k \qquad (6.52)$$

where x^p is the predicted value of x, $\hat{x}$ is the current estimate of x, $\hat{\dot{x}}$ denotes the current estimate of the rate of change of x, z is the noise contaminated measurement of x, T is the sample interval in seconds, and α and β are fixed real constants. If $\hat{\dot{x}} = 0$ for all k then this set of equations reduces to the first order filter, (6.40).

Cadzow [12] presents an analysis of the $\alpha\beta$ tracker in terms of the system transfer functions, and considers the specific case where a critically damped system is required. This constraint yields a relationship between α and β of the form

$$\alpha = 2\sqrt{\beta} - \beta \tag{6.53}$$

where β must be selected such that $0 < \beta < 4$. Values of β close to one yield a fast responding system, but values close to zero produce a system which is very sluggish.

If the $\alpha\beta$ tracker is constrained further to operate on a noise contaminated signal of the form (6.1) then the algorithm will generate an unbiased consistent estimator with an output variance given by

$$\sigma^2(\hat{x}) = K\,\sigma^2(z) \tag{6.54}$$

where, for a critically damped system,

$$K = \frac{10\sqrt{\beta} - 14\beta + 5\beta\sqrt{\beta}}{(2 - \sqrt{\beta})^3} \tag{6.55}$$

This function decreases as β approaches zero, thus enhancing the variance reduction, but increases rapidly beyond $\beta = 2$ causing the tracker performance to deteriorate rapidly.

In general, when selecting the values of α and β, a compromise exists between the conflicting requirements of good noise filtering (producing a sluggish system with a long time constant and narrow bandwidth) and of good transient following capability (resulting in a fast system with a short time constant and wide bandwidth). Benedict and Bordner [13] advocate the construction of a filter which will give the "best compromise" between the two extremes. Their solution results in a slightly underdamped system for which

$$\beta = \alpha^2/(2-\alpha) \tag{6.56}$$

They claim that by reducing the $\alpha\beta$ tracker to a form in which a single parameter, α, must be selected using the relationship (6.56), then the optimum estimates of both position and velocity are obtained simultaneously.

Although the $\alpha\beta$ tracker, in the optimised form presented above, results in the optimum linear fixed parameter tracker, for adaptive tracking it is suggested that α is allowed to vary with the observed high frequency power fluctuations of the error signal, $x_k - x_k^p$ [13]. A possible method for estimating α as a time dependent quantity could proceed by regarding the equation for $\hat{x}_{k+1}$ as forming a weighted average of the predicted value, x_{k+1}^p, and the measured value z_{k+1}; this becomes obvious if (6.50) is rearranged so that

$$\hat{x}_{k+1} = (1-\alpha)\, x_{k+1}^p + \alpha z_{k+1} \tag{6.57}$$

The minimum mean square estimate of this average is obtained if [ref. 2, p. 6],

$$\alpha = \frac{\mathrm{Var}(x^p)}{\mathrm{Var}(x^p) + \mathrm{Var}(z)} \tag{6.58}$$

but the problem then is to know what $\mathrm{Var}(x^p)$ and $\mathrm{Var}(z)$ are.

It has been suggested that a running average or first order recursive filter could be applied to update these quantities, or some other windowing filter could be used having a relatively short memory to ensure that α changes fairly rapidly in response to any loss of tracking [14]. An alternative is to use the differential quantities $\mathrm{Var}(z - x^p)$, and $\mathrm{Var}(\hat{x} - x^p)$ to determine α rather than $\mathrm{Var}(z)$ and $\mathrm{Var}(x^p)$ directly. However, if z and x^p or $\hat{x}$ and x^p, are partially correlated then

$$\mathrm{Var}(z - x^p) \lesssim \mathrm{Var}(z) + \mathrm{Var}(x^p)$$

$$\mathrm{Var}(\hat{x} - x^p) \lesssim \mathrm{Var}(\hat{x}) + \mathrm{Var}(x^p)$$

and the consequences of using these quantities would be unpredictable.

6.2.3 Recursive Least Squares

Young [9] provides a comprehensive treatment of recursive least squares analysis which commences with the relatively simple example for a single unknown parameter in a "regression relationship", and progresses through more complex cases which introduce probabilistic considerations and multi-parameter estimation. The technique of introducing extra information, of a probabilistic nature, has already been considered briefly in the discussion of Section 6.1.5 on Maximum Likelihood and Bayesian Estimation. Other than these two examples, however, very little has been assumed concerning the statistical properties of the signal or measurement noise formulations of either the recursive or non-recursive estimators considered so far. This absence of "a priori" statistical information has led to such estimators being referred to as deterministic.

The recursive least squares example quoted here represents an intermediate state in the a priori knowledge, between the relatively primitive estimators discussed previously and the Kalman filter which is capable of incorporating complex system and measurement models. In Young's treatment, the basic problem is the estimation of a set of unknown parameters a_j; $j = 1, 2, \ldots, n$, which appear in a linear "regression" relationship of the form

$$x_o = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \qquad (6.59)$$

where x_j; j = 1, 2, ..., n are exactly known, linearly independent variables. In vector form (6.59) becomes

$$x_o = \underline{x}^T \underline{a} \tag{6.60}$$

(the vector inner product) where T denotes the vector/matrix transpose.

In general, the observation, z_i, of x_{o_i} in the presence of measurement noise, v_i, then becomes,

$$z_i = \underline{x}_i^T \underline{a} + v_i;\ i = 1, 2, \ldots, k \tag{6.61}$$

Three basic assumptions are now made about the nature of the measurement noise, v_i;

(1) $E[v_i] = 0$

(2) $E[v_i\ v_j] = \sigma^2 \delta_{ij}$ where δ is the Kronecker delta symbol

(3) $E[\underline{x}_i\ v_j] = 0$ for all i,j

These simply state that the measurement noise is white (sequentially/temporally uncorrelated) of zero mean, and that all noise samples are uncorrelated with the independent variables, x_i, which compose the vector $\underline{x}_j$.

Assumptions (1) and (3) are important in demonstrating that the estimate $\hat{\underline{a}}_k$ is unbiased. Assumption (2), however, is used to generate a "covariance matrix" of the estimation errors, P_k, which has the form,

$$P_k = E[\tilde{\underline{a}}_k\ \tilde{\underline{a}}_k] = \begin{bmatrix} E[\tilde{a}_1], & E[\tilde{a}_1\ \tilde{a}_2], & \ldots, & E[\tilde{a}_1\ \tilde{a}_n] \\ E[\tilde{a}_2\ \tilde{a}_1], & E[\tilde{a}_2], & \ldots, & E[\tilde{a}_2\ \tilde{a}_n] \\ \vdots & \vdots & & \vdots \\ E[\tilde{a}_n\ \tilde{a}_1], & E[\tilde{a}_n\ \tilde{a}_2], & & E[\tilde{a}_n\ \tilde{a}_n] \end{bmatrix} \tag{6.62}$$

where $\underset{\sim}{\tilde{a}} = \hat{a}_k - \underset{\sim}{a}$ and $E[\underset{\sim}{\hat{a}}] = 0$.

The recursive least squares algorithm derived by Young is reproduced here:

$$\underset{\sim}{\hat{a}}_k = \underset{\sim}{\hat{a}}_{k-1} - P_{k-1}\underset{\sim}{x}_k\left[\hat{\sigma}^2 + \underset{\sim}{x}_k^T P_{k-1}\underset{\sim}{x}_k\right]^{-1}\left\{\underset{\sim}{x}_k^T\underset{\sim}{\hat{a}}_{k-1} - z_k\right\} \tag{6.63}$$

or

$$\underset{\sim}{\hat{a}}_k = \underset{\sim}{\hat{a}}_{k-1} - \frac{P_k}{\hat{\sigma}^2}\left\{\underset{\sim}{x}_k\underset{\sim}{x}_k^T\underset{\sim}{\hat{a}}_{k-1} - \underset{\sim}{x}_k z_k\right\} \tag{6.64}$$

and

$$P_k = P_{k-1} - P_{k-1}\underset{\sim}{x}_k\left[\hat{\sigma}^2 + \underset{\sim}{x}_k^T P_{k-1}\underset{\sim}{x}_k\right]^{-1}\underset{\sim}{x}_k^T P_{k-1} \tag{6.65}$$

The most significant difference between this and other estimators considered so far is the inclusion, in the algorithms above, of a recursively updated estimate of the error covariance which is dependent on the variance estimate, $\hat{\sigma}^2$, of the measurement noise, v. This variance, $\hat{\sigma}^2$, will be an a priori estimate derived either from the data itself or from some alternative independent source. A variance estimator such as (6.39) could be used but this form is biased because it does not include a "degree of freedom" adjustment. Such an adjustment is, however, provided by the simple en bloc solution,

$$\hat{\sigma}^2 = \frac{1}{k-n}\ \underset{\sim}{v}^T\underset{\sim}{v} = \frac{1}{k-n}\sum_{i=1}^{k} v_i \tag{6.66}$$

where the sum of the squares of the residuals is divided by k-n rather than just k, as in (6.39), and n is the dimension of the vector $\underset{\sim}{x}$.

Expression (6.65) also demonstrates that the error covariance is a strictly decreasing function of sample size, so that the precision of the estimate increases as more data is utilised. This behaviour is one manifestation of the statistical property of consistency.

As a final note on the least squares technique, in general, Young points out that this method of estimation can be interpreted from both

the maximum likelihood and Bayesian standpoints. For example, if it is further assumed that, in addition to its white noise properties, $\underline{v}$ has a gaussian amplitude distribution then it is also possible to assume that the estimation error $\underset{\sim}{\tilde{a}}$ also has the same distribution. Since the statistical characterisation of the gaussian distribution are completely determined by the first two statistical moments (in this case, the mean $\underline{\hat{a}}_k$ and the error covariance P_k), the recursive least squares algorithm can be regarded as a recursive maximum likelihood estimator [9].

If it is assumed that a priori information on the mean and covariance matrix is available in the form of the initial estimates $\underline{\hat{a}}_o$ and P_o, then the algorithm can be considered as a Bayesian estimator. This apparently is a consequence of the application of the "Bayes Rule" linking a priori and a posteriori probability statements [9].

6.3 Optimal Estimation

This section examines first the linear, discrete form of the Kalman filter and considers, briefly, the corresponding equations for the continuous form. Gelb's treatment [2] is adhered to closely in the presentation of Section 6.3.1. Other topics then address the problem of identifying parameters modelled into the filter, and specific implementations of the Kalman algorithm (selected from research literature) which are relevant to the processing of lidar data. Finally, methods for simulating lidar data are considered, together with comments on their relevance as an aid in evaluating filter performance.

6.3.1 The Kalman Filter

Although early work in control and estimation theory used frequency domain analysis, Kalman's results employ a time domain des-

scription formulated in "state space" notation. This notation has the dual advantages of being both less unwieldy and closer to physical reality than frequency domain analysis, and is particularly useful in providing a statistical description of system behaviour. The dynamics of a linear system can be described using either continuous or discrete forms of this notation.

In continuous form, such a system may be represented by the first order vector-matrix differential equation

$$\dot{\underline{x}}(t) = F(t)\underline{x}(t) + G(t)\underline{w}(t) + L(t)\underline{u}(t) \tag{6.67}$$

where $\underline{x}(t)$ is the system state vector, $\underline{w}(t)$ is a random forcing function, $\underline{u}(t)$ is a deterministic control input, and F(t), G(t), L(t) are matrices arising in the formulation. In discrete notation, this linear differential equation becomes

$$\underline{x}_{k+1} = \Phi_k \underline{x}_k + \Gamma_k \underline{w}_k + \Lambda_k \underline{u}_k \tag{6.68}$$

where Φ, Γ and Λ are also matrices, corresponding in function to the matrices F, G and L above.

Either notation may be used but the discrete form is preferred here since the lidar data to be analysed forms a sequence of sampled measurements. An adequate system model is obtained if the deterministic control input term is dropped and the matrix Γ_k is assumed to be identical to the identity matrix, I. Expression (6.68) then reduces to

$$\underline{x}_k = \Phi_k \underline{x}_{k-1} + \underline{w}_{k-1} \tag{6.69}$$

This represents a system whose state at time t_k is denoted by $\underline{x}(t_k)$, or simply $\underline{x}_k$, where $\underline{w}_k$ is a zero mean white noise sequence having a covariance matrix, $Q_k = E[\underline{w}_k \underline{w}_k^T]$.

The measurement, as in the recursive least squares example, is assumed to be a linear combination of the system state variables and uncorrelated measurement noise, such that

$$\underline{z}_k = H_k \underline{x}_k + \underline{v}_k \tag{6.70}$$

where

$$\underline{z}_k = [z_1, z_2, \ldots, z_\ell]$$

is a vector of ℓ measurements, taken at time t_k and $\underline{v}_k$ is a vector of random noise quantities all having a zero mean and a covariance matrix, $R_k = E[\underline{v}_k \underline{v}_k^T]$. The measurement matrix, H_k, describes the linear combination of state variables which comprise $\underline{z}_k$ in the absence of noise.

Having defined the system and measurement models, a linear recursive estimator is now required, having the general form;

$$\hat{\underline{x}}_k(+) = K'_k \hat{\underline{x}}_k(-) + K_k z_k \tag{6.71}$$

This expression states that each updated estimate, $\hat{\underline{x}}_k(+)$, is a linear combination of the previous estimate, $\underline{x}_k(-)$, and the current measurement, $\underline{z}_k$, weighted in significance according to the time varying matrices, K'_k and K_k.

If the relationship between the state estimate and the state error, $\tilde{\underline{x}}$ (+ or -), is defined as

$$\hat{\underline{x}}_k(+) = \underline{x}_k + \tilde{\underline{x}}_k(+)$$

$$\hat{\underline{x}}_k(-) = \underline{x}_k + \tilde{\underline{x}}_k(-) \tag{6.72}$$

(where (-) and (+) are used to denote terms immediately before and immediately after a discrete measurement) then, by substituting (6.72)

and (6.70), the following result may be obtained;

$$\underset{\sim}{\tilde{x}}_k(+) = [K'_k + K_k H_k - I]\underset{\sim}{x}_k + K'\underset{\sim}{\tilde{x}}_k(-) + K_k \underset{\sim}{v}_k \tag{6.73}$$

Since one of the properties required of an optimal estimator is that it be unbiased, it is therefore necessary that

$$E[\underset{\sim}{\tilde{x}}_k(-)] = 0 \tag{6.74}$$

Also, by definition, $E[\underset{\sim}{v}_k] = 0$, so in order to obtain the result (6.73), the following equivalence must be true;

$$K'_k = I - K_k H_k \tag{6.75}$$

The estimator of (6.71) then becomes

$$\underset{\sim}{\hat{x}}_k(+) = (I - K_k H_k)\,\underset{\sim}{\hat{x}}_k(+) + K_k \underset{\sim}{z}_k \tag{6.76}$$

or, alternatively,

$$\underset{\sim}{\hat{x}}_k(+) = \underset{\sim}{\hat{x}}_k(-) + K[\underset{\sim}{z}_k - H_k \underset{\sim}{\hat{x}}_k(-)] \tag{6.77}$$

Error Covariance Update

The form of the error covariance matrix was defined earlier in expression (6.62) but will be re-defined, here, using the current notation, such that

$$P_k = E[\underset{\sim}{\tilde{x}}_k \underset{\sim}{\tilde{x}}^T] \tag{6.78}$$

In the recursive least squares algorithm, only a single expression, (6.65), was required to update the error covariance upon receipt of the latest measurement. For the Kalman filter, however, the inclusion of a system and measurement model necessitates the derivation of both an "error

covariance update" equation and an "error covariance propagation/ extrapolation" equation. If the former is considered first, the appropriate form for (6.77) becomes

$$P_k(+) = E[\underset{\sim}{\hat{x}}_k(+)\ \underset{\sim}{\hat{x}}_k(+)^T] \tag{6.79}$$

Using equations (6.70), (6.72) and (6.77) the estimation error immediately after the receipt of a measurement may be expressed as

$$\underset{\sim}{\hat{x}}_k(+) = (I - K_k H_k)\ \underset{\sim}{\hat{x}}(-) + K_k \underset{\sim}{v}_k \tag{6.80}$$

Substitution of this result into (6.79) yields

$$P_k(+) = E\{(I - K_k H_k)\ \underset{\sim}{\hat{x}}_k(-)[\underset{\sim}{\hat{x}}_k(-)^T(I - K_k H_k)^T + \underset{\sim}{v}_k^T K_k^T]$$

$$+ K_k \underset{\sim}{v}_k[\underset{\sim}{\hat{x}}_k(-)^T(I - K_k H_k)^T + \underset{\sim}{v}_k^T K_k]\} \tag{6.81}$$

However, since by definition

$$E[\underset{\sim}{\hat{x}}_k(-)\ \underset{\sim}{\hat{x}}_k(-)^T] = P_k(-) \tag{6.82}$$

and

$$E[\underset{\sim}{v}_k\ \underset{\sim}{v}_k^T] = \bar{R}_k \tag{6.83}$$

then, by incorporating the requirement that measurement errors be uncorrelated,

$$E[\underset{\sim}{\hat{x}}_k(-)\ \underset{\sim}{v}_k^T] = E[\underset{\sim}{v}_k\ \underset{\sim}{\hat{x}}_k(-)^T] = 0 \tag{6.84}$$

the resultant form for the error covariance update becomes

$$P_k(+) = (I - K_k H_k)\ P_k(-)(I - K_k H_k)^T + K_k R_k K_k^T \tag{6.85}$$

The next step is to optimise the gain, K_k, and this is accomplished by minimising the weighted scalar sum of the diagonal elements of the

error covariance matrix, $P_k(+)$. Using the same teechniques as employed in Section 6.1.4, a cost function of the form

$$J_k = E[\underset{\sim}{x}(+)^T \underset{\sim}{x}(+)] = \text{trace}[P_k(+)] \tag{6.86}$$

is obtained and its partial derivative with respect to K_k is set equal to zero. By substituting (6.85) into (6.86), the result obtained is

$$-2(I - K_k H_k)\, P_k(-)\, H_k^T + 2K_k R_k = 0 \tag{6.87}$$

or by solving for K_k,

$$K_k = P_k(-)\, H_k^T [H_k P_k(-)\, H_k^T + R_k]^{-1} \tag{6.88}$$

which is referred to as the "Kalman Gain Matrix". If (6.88) is now substituted into (6.85) a new form for $P_k(+)$ is obtained;

$$P_k(+) = [I - K_k H_k]\, P_k(-) \tag{6.89}$$

which is the optimised value for the updated estimation error covariance matrix.

State Transition Matrix Extrapolations

A distinction is generally made between those quantities which are updated "across" a measurement, such as the error covariance of expression (6.89) above, and the extrapolation of these quantities "between" measurements. It is possible to demonstrate [2] that an extrapolated but unbiased estimate of the state variable, and a projected value for the error covariance matrix, may be obtained over the interval between measurements using the state transition matrix, Φ. The appropriate relationships are

$$\hat{\underline{x}}_k(-) = \Phi_{k-1}\,\hat{\underline{x}}_{k-1}(+) \tag{6.90}$$

$$P_k(-) = \Phi_{k-1}P_{k-1}(+)\,\Phi^T + Q_{k-1} \tag{6.91}$$

Expressions (6.69), (6.70), (6.77) and (6.89) through to (6.91) form the essential components of the Kalman filter algorithm. If $\underline{w}_k$ and $\underline{v}_k$ are both gaussian random variables, this algorithm represents the optimum linear filter - a non-linear filter cannot do any better. Figure 6.4 illustrates the complete discrete system in a block diagram form which emphasises the functional separation between the mathematical abstraction of what the system and measurement processes are believed to be, and the filter itself. There is no feedback from the state equation, to those calculations at the covariance level, which ultimately serve to provide K_k The computer flow diagram of Figure 6.5 reflects this partitioning of the algorithm components.

The Continuous Kalman Filter

In continuous form, the system and measurement models corresponding to (6.69) and (6.70) are

$$\dot{\underline{x}} = F(t)x(t) + G(t)\underline{w}(t) \tag{6.92}$$

$$\underline{z} = H(t)\underline{x}(t) + \underline{v}(t) \tag{6.93}$$

where $\underline{w}$ and $\underline{v}$ are zero mean white noise processes with spectral density matrices Q and R, respectively. Gelb demonstrates the transition from the discrete to the continuous formulation using the following equivalences

$$\Phi_k \rightarrow I + F\Delta t$$

$$Q_k \rightarrow GQG^T\Delta t \tag{6.94}$$

$$R_k \rightarrow R/\Delta t$$

Without detailing the necessary steps in the transition between discrete and continuous systems, a summary of the important continuous Kalman filter equations is presented below.

The state estimate becomes

$$\dot{\underline{\hat{x}}}(t) = F(t)\underline{\hat{x}}(t) + K(t)[\underline{z}(t) - H(t)\underline{\hat{x}}(t)] \tag{6.95}$$

and the error covariance propagation is

$$\dot{P}(t) = F(t)P(t) + P(t)F^T(t) + G(t)Q(t)G^T(t) - K(t)R(t)K^T(t) \tag{6.96}$$

The Kalman Gain Matrix is then

$$K(t) = P(t)H^T(t)R^{-1}(t) \text{ when } E[\underline{w}(t)\ \underline{v}^T(\tau)] = 0 \tag{6.97}$$

or

$$K(t) = [P(t)H^T(t) + G(t)C(t)]T^{-1}(t) \tag{6.98}$$

when $E[\underline{w}(t)\ \underline{v}^T(\tau)] = C(t)\ \delta(t-\tau)$, where $C(t)$ is an autocovariance function.

Expression (6.96) is known as the "matrix Riccati equation" and is a useful complement to the discrete filter algorithms for predicting steady state error covariances, ($\dot{P} = 0$). Given the steady state error covariance, the steady state Kalman gain is then readily calculated using (6.97).

6.3.2 Model Identification (The Box-Jenkins Approach)

Although Box and Jenkins were not the first to study time series models, their names are associated with the subject due principally to its popularisation via their book, "Time Series Analysis: Forecasting and Control" [15]. A time series may be defined as a collection of observations, ordered with respect to time, which are usually expected to be dependent [16]. In general they represent an actual realisation of some underlying process, and the objective of time series analysis is to describe succinctly this theoretical process in the form of an observable model having similar properties to those of the process itself.

In order to make meaningful estimates of the primary statistical characteristics of a time series (mean, variance and autocorrelation function), from a single realisation, it is necessary that such a series exhibits the property of "stationarity". Stationarity, as it is defined in the "narrow sense", implies not only the absence of any trend, but also a mean and variance which are both constant and finite. Applied in the "wide sense", stationarity means that all statistical moments are constant and finite. A further requirement is that the autocorrelation between values of the process at two time periods depends only on the separation between these two time points and not on the absolute value of the time period itself. If the time series is denoted by z_t, these requirements may be summarised as follows:

$$\mu = E[z_t] \tag{6.99}$$

$$\sigma^2 = E[(z_t - \mu)^2] \tag{6.100}$$

and

$$\rho(z_t, z_s) = E[(z_t - \mu)(z_s - \mu)]/\sigma^2 \qquad t > s \tag{6.101}$$

where μ, σ^2 and $\rho(z_t, z_s)$ are the mean, variance and autocorrelation respectively.

Since many time series, derived from real life processes, are not stationary, various techniques are recommended for inducing stationarity so that methods of statistical analysis may be properly applied. For example, least squares techniques can be applied to fit a polynomial to the data with the intention of approximating the trend, thereby making it possible to remove it. However, in practice, it is extremely difficult to decide whether a change in the level of a series is of deterministic or probabilistic origin.

Box and Jenkins therefore advocate the use of an alternative method called "differencing" which involves subtracting the observation from one another in some prescribed time-dependent order. First order differencing, for example, is defined as the difference between the values of two adjacent observations, and second order differencing then consists of taking the differences of a differenced series etc. If the non-stationarity is a property of the variance, other techniques exist for "stabilising" the variance; for example, if the variance is proportional to the mean level of a series, or changes at a constant percentage rate, then a logarithmic transformation may be employed.

The following outline of the properties of "Autoregressive, Moving Average" models was derived from Vandaele [16].

Autoregressive Models

A time series is said to be governed by a first order autoregressive (AR) process if the current value of the time series, z_t, can be expressed as a linear function of the previous value of the series, z_{t-1}, and a random shock, a_t. It may be written as

$$z_t = \phi_1 z_{t-1} + a_t \tag{6.102}$$

where ϕ_1 describes the effect of a unit change in z_{t-1} on z_t and a_t is a zero mean, white noise series of constant variance, σ_a^2, such that

$$E[a_t] = 0 \tag{6.103}$$

$$E[a_t\ a_s] = \begin{cases} \sigma_a & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases} \tag{6.104}$$

and $$E[a_t, z_{t-1}] = 9 \tag{6.105}$$

If (6.102) is used to model a real data series, it will be necessary to estimate the value of the autoregressive parameter ϕ_1.

In general, the z_t's are regarded as deviations from the mean, μ, so that (6.102) could also be presented as

$$(y_t - \mu) = \phi_1(y_{t-1} - \mu) + a_t \tag{6.106}$$

or

$$y_t = (1 - \phi_1)\mu + \phi_1 y_{t-1} + a_t \tag{6.107}$$

where μ is the mean of y_t and y_t represents the actual data. Such a process is referred to as an autoregressive process of order 1, and denoted AR(1), where the order corresponds to the number of parameters, ϕ, that need to be estimated.

Taking the expectation value of (6.102);

$$\begin{aligned} E[z_t^2] &= E[(\phi_1 z_{t-1} + a_t)^2] \\ &= E[(\phi_1 z_{t-1})^2 + 2\phi_1 z_{t-1} a_t + a_t^2] \\ &= \phi_1 E[z_{t-1}^2] + 2\phi E[z_{t-1}\ a_t] + E[a_t^2] \end{aligned} \tag{6.108}$$

However, since the z_t's represent deviations from the mean, μ, of a stationary series then, by definition,

$$E[z_t^2] = \sigma_z^2 = E[z_{t-1}^2] \tag{6.109}$$

Also, the second term on the right hand side disappears (using the expression (6.105)), and the third term is simply the variance of a_t (according to (6.104)) so that (6.108) becomes

$$\sigma_z^2 = \frac{\sigma_a^2}{1 - \phi_1^2} \tag{6.110}$$

For finite values of the variance, σ_z^2, ϕ_1 is, therefore, limited to values of

$$\phi_1^2 < 1 \text{ or } |\phi_1| < 1 \tag{6.111}$$

The autocorrelation function of an AR(1) process may be obtained by considering, first, the autocovariance of z_t at lag = 1, which is defined as

$$\lambda_1 = \text{Cov}(z_t, z_{t-1}) = E[z_t \, z_{t-1}] \tag{6.112}$$

By substituting (6.102) into (6.112), it is found that

$$\lambda_1 = \phi_1 \, E[z_{t-1}^2] + E[a_t \, z_{t-1}]$$

$$= \phi_1 \lambda_o = \phi_1 \, \sigma_z^2 \tag{6.113}$$

since $e[a_t \, z_{t-1}] = 0$. This process may be repeated for lag = 2, again substituting for z_t using (6.102), and the result obtained will be

$$\lambda_2 = \phi_1^2 \lambda_o \tag{6.114}$$

Proceeding in the same way for all $k > 0$, the general result obtained is

$$\lambda_k = \phi_1^k \lambda_o \tag{6.115}$$

so that the autocorrelation becomes

$$\rho_k = \frac{\lambda_k}{\lambda_o} = \phi_1^k \text{ for } k > 0 \tag{6.116}$$

Clearly, then, for values of $|\phi_1| < 1$, the autocorrelations of an AR(1) process will appear to decay exponentially to zero at a rate governed by the magnitude of ϕ_1.

Higher order AR models exist, having more than one lagged variable, which may be written as

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-1} + \ldots + \phi_p z_{t-p} + a_t \tag{6.117}$$

or simply denoted as an AR(p) process. They, of course, have much more complicated autocovariance functions than (6.115) and, in general, AR process models with $p > 2$ are rarely invoked because of their complexity.

Moving Average Models

Since the basic AR(1) model is obtained simply by summing a lagged value of z with a single noise variable, a, it is perhaps to be expected that another type of series may be generated by omitting the lagged variable, z, but including a lagged noise term, a_{t-1}. The resulting series,

$$z_t = a_t - \theta_1 a_{t-1} \tag{6.118}$$

is known as a first order "Moving Average" process, or MA(1). Again, in terms of the actual data, as opposed to deviations from the mean, μ, (6.118) becomes

$$\gamma_t - \mu = a_t - \theta_1 a_{t-1}$$

or

$$\gamma_t = \mu + a_t - \theta_1 a_{t-1} \tag{6.119}$$

It is possible to demonstrate that one advantage of the MA model is its ability to represent a process with an infinite number of autoregressive terms [16]. For example, the MA(1) process will generate the series

$$z_t = - \theta_1 z_{t-1} - \theta_1^2 z_{t-1} - \theta_1^3 z_{t-3} \ldots + a_t \tag{6.120}$$

and is therefore an economical way of representing complicated models, even though the autoregressive parameters are just powers of the same basic coefficient, θ_1.

Using, again, the white noise conditions specified in (6.103), (6.104) and (6.105), the MA(1) process has a variance derived as follows;

$$E[z_t] = E[a_t - \theta_1 a_{t-1}] = E[a_t] - \theta_1 E[a_{t-1}] = 0 \tag{6.121}$$

$$\begin{aligned} \sigma_z^2 &= E[(a_t - \theta_1 a_{t-1})^2] \\ &= E[(a_t^2 - 2\theta_1 a_{t-1} a_t + \theta_1^2 a_{t-1}^2] \\ &= (1 + \theta_1^2)\, \sigma_a^2 = \lambda_o \end{aligned} \tag{6.122}$$

The covariance at lag = 1 then becomes

$$\begin{aligned} \mathrm{Cov}(z_t, z_{t-1}) &= E[(a_t - \theta_t a_{t-1})(a_{t-1} - \theta_1 a_{t-2})] \\ &= E[- \theta_1 a_{t-1}^2] \\ &= - \theta_1 \sigma_a^2 \end{aligned} \tag{6.123}$$

However, repeating the calculation for lag = 2, gives

$$\lambda_2 = \text{Cov}(z_t, z_{t-2}) = E[(a_t - \theta_1 a_{t-1})(a_{t-2} - \theta_1 a_{t-3})]$$

$$= 0 \qquad (6.124)$$

due to the condition, (6.104). It can be similarly demonstrated that $\lambda_k = 0$ for all $k > 2$. The autocorrelations therefore become

$$\rho_1 = -\theta_1/(1 + \theta_1^2)$$

$$\rho_k = 0, \quad k \geq 2 \qquad (6.125)$$

Moving average models of some higher order, 1, can be expressed as

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \qquad (6.126)$$

or, simply as MA(q).

ARMA and ARIMA Models

An AR(1) and an MA(1) process may be combined into a single model referred to as an ARMA(1,1) which has the form

$$z_t = \phi_1 z_{t-1} + a_t - \theta_1 a_{t-1} \qquad (6.127)$$

It can be shown [16] that this model has a variance given by

$$\sigma_z^2 = \lambda_o = \frac{1 + \theta_1^2 - 2\theta_1\phi_1}{1 - \phi_1^2} \sigma_a^2 \qquad (6.128)$$

with an autocovariance at lag = 1, of

$$\lambda_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 - \phi_1^2} \sigma_a^2 \qquad (6.129)$$

and for higher lags,

$$\lambda_k = \phi_1 \lambda_{k-1} = \phi_1^{k-1} \lambda_1 \tag{6.130}$$

The autocorrelation terms then become

$$\rho_1 = \frac{\lambda_1}{\lambda_o} = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 + \theta_1^2 - 2\phi_1\theta_1}$$

$$\rho_k = \frac{\lambda_k}{\lambda_o} = \phi_1 \rho_{k-1}, \quad k \geq 2 \tag{6.131}$$

Higher order ARMA models are represented by a series of the form

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \tag{6.132}$$

which may be abbreviated to ARMA(p,q).

The technique of differencing was presented earlier as the method recommended by Box and Jenkins for rendering a series stationary. If it is now assumed, for example, that some arbitrary series, z_t, is non-stationary and that the first difference of the series generates the stationary series w_t, where

$$w_t = z_t - z_{t-1} \tag{6.133}$$

then z_t and z_{t-1} could be replaced in the ARMA(1,1) model of (6.127) to produce

$$w_t = \phi_1 w_{t-1} + a_t - \theta_1 a_{t-1} \tag{6.134}$$

This differenced series is then of the ARMA(1,1) form but the process for z_t is referred to as an "Autoregressive, Integrated Moving Average", or an ARIMA(1,1,1) model. As a model of some arbitrary order, this

becomes an ARIMA(p,d,q) process where d indicates the degree of differencing.

Model Identification Using Sampled Processes

Each of the models discussed above and presented in order of increasing complexity possess different autocorrelation functions (acfs), representative examples of which are illustrated in Figure 7.9. These assist in the identification of suitable models which may be regarded as at least an approximation to those processes observed in real data samples. Many processes, however, possess very similar autocorrelation functions. Estimates of these autocorrelation functions obtained from real, and therefore finite, data sequences are known as "sample autocorrelation functions" and are subject to estimation errors which further compound the difficulties of uniquely identifying a process. The sample autocorrelation functions, r_k, may be obtained simply by applying the formula

$$r_k = \frac{c_k}{c_o}$$

where c_k is defined as

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} z_t z_{t+k} \qquad k \geq 0$$

and n is the total number of measurements, z, within a sample.

Another statistic which may be employed, however, is the "partial autocorrelation function " (pacf). It can be shown that any ARIMA(p,d,q) model can always be expressed as a pure autoregressive model possessing an acf which, although it may decay quickly, could also stretch out to infinity [16]. The partial autocorrelation function constitutes a

device for summarising all the information contained in the acf of an AR process in a small number of non-zero statistics. Unlike autocorrelations, however, pacfs cannot be estimated using a simple formula but are calculated from a solution of a set of equations known as the "Yule-Walker" equation system [16].

In any AR model of order, k, successfully applied as a model for real data, the coefficient ϕ_k represents a measure of the "excess" correlation not accounted for by an AR model of order (k-1). This, highest order, autoregressive coefficient (in this case ϕ_k) is defined as the partial autocorrelation at lag k and is denoted by ϕ_{kk}. Therefore, as an example, if a time series process is actually an AR(k-1) process, then by successively calculating whether the parameters ϕ_1, ϕ_2, ..., ϕ_{k-1}, ϕ_k should be included in the model, it would be found that ϕ_{11}, ϕ_{22}, ..., $\phi_{k-1,k-1}$ would all have values which are different from zero, but that the value of ϕ_{kk} would be zero.

Box and Jenkins outline a method for calculating both the acf and pacf of a time series by applying the techniques of "Univariate Stochastic Model Identification" (USID). Details of this method, in a form intended for computer implementation, are appended to the main text of their book [ref. 15, part V, algorithm 1] and are to be applied in the data analysis of Chapter 7.

6.3.3 Estimating System and Measurement Noise Parameters

Application of the Kalman filter algorithm, presented in Section 6.3.1, to linear dynamic systems of the kind discussed in Section 6.3.2 require an exact knowledge of the process noise covariance matrix, Q, and the measurement noise covariance matrix, R. Often, however, both

Q and R are either unknown or are known only approximately. The investigation of methods for identifying Q and R forms a significant area of research in its own right, which is particularly active on the topic of "adaptive filtering". This refers to the real time estimation of parameters such as Q and R "on line", with consequent modifications to the error covariance and Kalman gain matrices, necessary to track the non-stationarity inherent in many real signal and measurement noise processes.

The technique discussed here and applied in the analysis of Chapter 7 originated with R.K. Mehra in a paper entitled "On the Identification of Variances and Adaptive Kalman Filtering" [17]. His analysis is limited to the case in which the statistical properties of Q and R must be regarded as stationary, but other techniques have since been published which are claimed to be capable of handling the more general time varying case [17], or are regarded as superior in some other respect [19,20]. Not all provide estimates of Q and R; some estimate the optimised Kalman gain directly. Many of these, however, appear to have been stimulated by Mehra's original contribution.

As an investigation into the relative merits of those different algorithms was considered beyond the scope of present work, Mehra's technique is used in the analysis of Chapter 7, since it generally produced satisfactory results. It is acknowledged, however, that some of the other algorithms may eventually prove to be superior for the purposes of processing lidar data.

The Mehra Algorithm for Estimating Q and R

It is first assumed that the system under consideration is time invariant and that both the system and filter (optimal or sub-optimal)

have reached steady state conditions (the latter corresponds to $\dot{P} = 0$ in (6.96)). Both the system and filter equations are identical to those presented in Section 6.3.1 but Mehra uses K_o to identify the suboptimal steady state Kalman filter gain,

$$K_o = M_o H^T (H M_o H^T + R_o)^{-1} \tag{6.135}$$

where

$$M_o = \Phi[M_o - M_o H^T (H M_o H^T + R_o)^{-1} H M_o]\Phi^T + \Gamma Q_o \Gamma^T \tag{6.136}$$

is the steady state solution to the error covariance equation of (6.85) and Q_o and R_o denote initial estimates of Q and R.

In an optimal Kalman filter, for which $Q_o = Q$ and $R_o = R$, M_o is the optimised error covariance. However, in the sub-optimal case, the error covariance estimate becomes [17]

$$M_1 = \Phi[M_1 - K_o H M_1 - M_1 H^T K_o + K_o (H M_1 H^T + R) K_o^T]\Phi^T + \Gamma Q \Gamma^T \tag{6.137}$$

where

$$M_1 = E[(x_i - \hat{x}_i(-))(x_i - \hat{x}_i(-))] \tag{6.138}$$

To check whether the Kalman filter constructed using some estimate of Q and R is close to the optimum or not, the statistical technique of "hypothesis testing" is used with the autocorrelation function of the innovation sequence.

For an optimal filter, the innovation sequence,

$$\nu_i = z_i - H\hat{x}_i(-) \tag{6.139}$$

of the filter equation (6.80) is a Gaussian white noise sequence. This becomes apparent if the error in the state estimate

$$e_i = x_i - \hat{x}_i(-) \tag{6.140}$$

and the measurement equation (6.70) are substituted into (6.139). The result obtained is,

$$\begin{aligned} \nu_i &= z_i - H\hat{x}(-) \\ &= Hx_i + v_i - Hx_i(-) \\ &= H(x_i - \hat{x}_i(-)) + v_i \\ &= He_i + v_i \end{aligned} \tag{6.141}$$

and by forming the autocovariance function of ν_i,

$$\begin{aligned} E[\nu_i \nu_i^T] &= E[(He_i + v_i)(He_j + v_j)^T] \\ &= 0 \text{ for } j \neq i \end{aligned} \tag{6.142}$$

since v_i is independent of both e_j and v_j, for $i > j$, and v_j is independent of e_i and v_i, for $i < j$. Also, Mehra states that

$$E[\nu_i \nu_j] = HMH^T + R \tag{6.143}$$

for $i = j$, and since ν_i is a linear sum of Gaussian random variables, it is also Gaussian.

It is further demonstrated by Mehra, that the autocorrelation function of ν_i does not depend on i, so that ν_i is a stationary Gaussian random sequence. If the autocovariance of the innovation sequence is now given the notation,

$$C_k = E[\nu_i \nu_{i-k}^T]$$

then [17],

$$C_k = HMH^T + R \qquad k = 0 \qquad (6.144)$$

$$= H[\Phi(I - KH)]^{k-1}\Phi[MH^T - KC_o] \qquad k > 0 \qquad (6.145)$$

Upon substitution of the optimum choice of K (expression (6.88) with M replacing P), C_k vanishes for all $k > 0$.

In order to test the innovation sequence, an estimate of C_k is required which is obtained using the relationship

$$\hat{C}_k = \frac{1}{N}\sum_{i=k}^{N} \nu_i \nu_{i-k}^T \qquad (6.146)$$

Dividing by N instead of (N-k) yields a biased estimate but Mehra argues that (6.146) is preferable since it gives less mean square error than the unbiased form. Using approximate results, valid in the limit of large N, he also shows that for an innovation sequence consisting of white noise,

$$\mathrm{Var}[\hat{\rho}_k] = \mathrm{Var}\left[\frac{\hat{C}_k}{C_o}\right] = \frac{1}{N} \qquad (6.147)$$

Standard statistical procedures are then applied; having established that the 95% confidence limits for $\{\hat{\rho}_k\}$, $k > 0$, are $\pm(1.96/N^{\frac{1}{2}})$ the set of values $\{\hat{\rho}_k\}$ are examined and the number of times they lie outside the band $\pm(1.96/N^{\frac{1}{2}})$ is counted. If this number is less than 5% of the total, the sequence ν_i is said to be white.

For the analysis of Chapter 7, the set of autocorrelations $\{\hat{\rho}_k\}$ contained 20 values (k = 1, 20). Since Mehra's whiteness test is ultimately derived from a simple statistical test based on confidence limits it is, of course, not really necessary to discriminate abruptly between white and non-white innovation sequences. Clearly, however, as the number of values falling outside the confidence limits decreases towards zero,

better filter performance can be expected.

Estimating Q and R

If the above test reveals that the filter is sub-optimal, better estimates of Q and R are required. These estimates are derived from the innovation sequence autocorrelation function and the method proceeds in three steps [17]:

(1) First an estimate of the matrix product, MH^T is required which is obtained via the expression (6.145). If this relationship is expanded into its component terms such that,

$$C_1 = H\Phi MH^T - H\Phi KC_o$$

$$C_2 = H\Phi^2 MH^T - H\Phi KC_1 - H\Phi^2 KC_o$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$C_n = H\Phi^n MH^T - H\Phi KC_{n-1} - \ldots - H\Phi^n KC_o$$

where n is the state vector dimension, then it is possible to show that [17],

$$MH^T = B^* \begin{bmatrix} C_1 + H\Phi KC_o \\ C_2 + H\Phi KC_1 + H\Phi^2 KC_o \\ \vdots \qquad \vdots \qquad \vdots \\ C_n + H\Phi KC_{n-1} + \ldots + H\Phi^n KC_o \end{bmatrix} \tag{6.148}$$

where B^* is the pseudo-inverse of matrix B which is defined as

$$B = \begin{bmatrix} H \\ H\Phi \\ \vdots \\ H\Phi^{n-1} \end{bmatrix} .\Phi \qquad (6.149)$$

and $B^* = (B^T B)^{-1} B^T$.

By substituting estimated autocorrelation terms into (6.148), an estimate of MH^T may be obtained;

$$\hat{M}\hat{H}^T = \begin{bmatrix} \hat{C}_1 + H\Phi K\hat{C}_o \\ \hat{C}_2 + H\Phi K\hat{C}_1 + H\Phi^2 K\hat{C}_o \\ \vdots \qquad \vdots \qquad \vdots \\ \hat{C}_n + H\Phi K\hat{C}_{n-1} + \ldots + H\Phi^n K\hat{C}_o \end{bmatrix} \qquad (6.150)$$

Mehra, however, gives an alternative form for this estimate, also derived directly from (6.145);

$$\hat{M}\hat{H}^T = K\hat{C}_o + A^* \begin{bmatrix} \hat{C}_1 \\ \vdots \\ \hat{C}_r \end{bmatrix} \qquad (6.151)$$

where

$$A = \begin{bmatrix} H\Phi \\ H\Phi(I - KH)\Phi \\ \vdots \\ H[\Phi(I - KH)]^{n-1}\Phi \end{bmatrix}$$

which, as a result of experimental observations, he has found preferable to use since matrix A is better conditioned than matrix B.

(2) Having obtained the above estimate of MH^T, an estimate of R can now be derived, simply by using (6.144);

$$\hat{R} = \hat{C}_o - H(\hat{M}\hat{H}^T) \tag{6.152}$$

(3) Finally, an estimate of Q must be obtained. A complication arises here due to the fact that, in general, the state vector dimension, n, need not be identical to the dimension of the measurement vector, r (where $\underline{x}$ and $\underline{z}$ are used to denote the state vector and measurement vector respectively - see expressions (6.69) and (6.79)). This means that although the steady state error covariance matrix, M, is square and of dimension nxn, the measurement matrix H has the dimensions rxn. Consequently, the estimated quantity, $\hat{M}\hat{H}^T$, above will have the dimensions, nxr, which means that only nxr linear relationships between the unknown elements of Q are available [17]. Therefore if the number of unknowns in Q is nxr or less, a solution can be obtained, but if the number of unknowns is larger than this, then a unique solution does not exist. Fortunately, for the scalar case examined in Chapter 7, n = r = 1, so a unique solution for Q = q can be obtained.

If the expression for the sub-optimal error covariance estimate is represented in terms of the steady state error covariance (optimal or sub-optimal), then

$$M = \Phi(I - KH)M(I - KH)^T\Phi^T + \Phi KRK^T\Phi^T + \Gamma Q\Gamma^T \tag{6.153}$$

which Mehra solves for the unknown elements of Q by re-writing as

$$M = \Phi M\Phi^T + \Omega + \Gamma Q\Gamma^T \tag{6.154}$$

where

$$\Omega = \Phi[- KHM - MH^TK^T + KC_oK^T]\Phi^T \tag{6.155}$$

Substituting back for M on the right hand side of (6.154) yields

$$M = \Phi^2M(\Phi^2)^T + \Phi\Omega\Phi^T + \Omega + \Phi\Gamma Q\Gamma^T\Phi^T + \Gamma Q\Gamma^T \tag{6.156}$$

then, by repeating the procedure n times, and separating the terms involving Q on the left hand side of the equation, Mehra obtains the result,

$$\sum_{j=0}^{k-1} \Phi^j\Gamma Q\Gamma^T(\Phi^j)^T = M - \Phi^kM(\Phi^k)^T - \sum_{j=0}^{k-1} \Phi^j\Omega(\Phi^j)^T \tag{6.157}$$

$$\text{for } k = 1, 2, \ldots, n$$

By pre-multiplying both sides of (6.157) by H and post-multiplying by

$$(\Phi^{-k})^T H^T$$

the following result is obtained,

$$\sum_{j=0}^{k-1} H\Phi^j\Gamma\hat{Q}\Gamma^T(\Phi^{j-k})^TH^T = \hat{M}\hat{H}(\Phi^{-k})^TH^T - H\Phi^k\hat{M}\hat{H}^T - \sum_{j=0}^{k-1} H\Phi^j\hat{\Omega}(\Phi^{j-k})^TH^T \tag{6.158}$$

$$k = 1, \ldots, n$$

where estimated values have been substituted for the theoretical values, and where

$$\hat{\Omega} = \Phi[- K\hat{H}\hat{M} - \hat{M}\hat{H}^TK^T + K\hat{C}_oK^T]\Phi^T \tag{6.159}$$

The right hand side of (6.158) is completely determined by $\hat{M}\hat{H}^T$ and $\hat{C}_o$, but the set of equations this relationship generates is not linearly independent. For any particular, non-scalar, case it is necessary to choose a linearly independent subset of these equations.

Mehra proceeds from this point to demonstrate how the optimal gain

may be estimated directly. Direct estimation of this gain has the advantage that even if the number of unknowns in Q is greater than nxr, a unique solution for the optimum gain, alone, may still be obtained. In the analysis of Chapter 7, however, estimates of Q and R are obtained in preference to the optimised gain, K_{op}, because of their relevance to the a priori requirement of system and measurement model identification.

6.3.4 Estimation of Sampled Data Containing Sequentially Correlated Noise

An important limitation of the basic linear, discrete Kalman filtering algorithm of Section 6.3.1 is that the measurement process of (6.70) is restricted to sequentially uncorrelated noise. Upon examination of the Hull lidar data, however, using the methods of Section 6.3.2, it becomes apparent that a correlated measurement noise process was probably making a significant contribution to the autocorrelation function of the data. Similar observations had been reported by at least one other research group involved in CO_2 lidar [1]. For the data examined in Chapter 7 this process appears, generally, to be restricted to approximately the first ten lags, implying the existence of correlation over periods of the order of 1 second, assuming a 10 Hz data capture rate.

Sampling of the data at intervals equal to or exceeding 10 lags would presumably remove this correlation problem but the existence of correlation in the measurement noise implies that an upper limit exists to the rate at which independent data samples may be collected. The modification to the basic Kalman filtering algorithm presented here is attributable to Bryson and Hendrikson [21] and describes a technique which is potentially capable of coping with temporally correlated

measurement noise.

A concise statement of the problem may be made by re-presenting the system state and measurement equations using Bryson and Hendrikson's notation;

$$\left.\begin{aligned} \text{State:}\quad & x_{i+1} = \Phi x_i + w_i \\ \text{Measurement:}\quad & z_i = Hx_i + \varepsilon_i \\ & \varepsilon_{i+1} = \Psi \varepsilon_i + u_i \end{aligned}\right\} \tag{6.160}$$

where w_i and u_i are Gaussian random vector sequences with zero means and covariances, Q and $\bar{Q}$, respectively. The state equation remains unchanged, but the measurement has two components, one of which describes the sequential correlation.

In general, Ψ, can be regarded as being time dependent, but for the application considered here, a constant coefficient system is assumed. A conventional approach to this problem involves "augmenting" the state vector so that the correlation parameter, Ψ, is recursively estimated together with the system process, x. State vector augmentation, however, also entails augmentation of the system matrices so that, in general, the augmented system has matrix components of the form,

$$x_i^a \equiv \begin{bmatrix} x_i \\ \hline \varepsilon_i \end{bmatrix}, \qquad H^a \equiv [H \,\vdots\, I]$$

$$\Phi^a \equiv \left[\begin{array}{c:c} \Phi & 0 \\ \hdashline 0 & \Psi \end{array}\right], \quad Q^a \equiv \left[\begin{array}{c:c} Q & 0 \\ \hdashline 0 & \bar{Q} \end{array}\right] \tag{6.161}$$

The system description would then become

$$x^a_{i+1} = \left[\begin{array}{c|c} \Phi & 0 \\ \hline 0 & \Psi \end{array}\right] x^a_i + \left[\begin{array}{c} w_i \\ \hline u_i \end{array}\right] \tag{6.162}$$

and the measurement is reduced to,

$$z_i = H^a x^a_i \tag{6.163}$$

so that the measurement noise matrix is zero ($R_k = 0$ and R_k^{-1} does not exist). Whilst this presents no problem as far as the Kalman gain matrix is concerned (see expression (6.88)), both Gelb [2] and Bryson and Hendrikson, point out that, if Q is small and $\Phi \simeq I$, then the error covariance update may become "ill-conditioned";

$$P_{k+1}(-) \simeq P_k(+) \tag{6.164}$$

Since both of these conditions pertain to the system model used in Chapter 7, an alternative to state vector augmentation is required.

In Bryson and Hendrikson's alternative a linear combination of z_{i+1} and z_i is used which does not contain ε_i;

$$\begin{aligned} \zeta_i &= z_{i+1} - \Psi z_i \\ &= Hx_{i+1} + \varepsilon_{i+1} - \Psi Hx_i - \Psi\varepsilon_i \\ &= Hx_{i+1} + \Psi\varepsilon_i + u_i - \Psi Hx_i - \Psi\varepsilon_i \\ &= H(\Phi x_i + w_i) + u_i - \Psi Hx_i \\ &= (H\Phi - \Psi H)x_i + Hw_i + u_i \end{aligned} \tag{6.165}$$

The transformed measurement, ζ_i, now contains only the purely random sequence, $Hw_i + u_i$ instead of the sequentially correlated sequence, ε_i. It is therefore convenient to re-state the problem as

$$\left.\begin{aligned} \text{state:}\quad & x_i = \Phi x_{i-1} + w_{i-1} \;;\quad w_{i-1} : (0,Q) \\ \text{measurement:}\quad & \zeta_{i-1} = H^r x_{i-1} + u_i + Hw_{i-1} \;;\; u_{i-1} : (0,\bar{Q}) \\ \text{where}\quad & H^r = H\Phi - \Psi H \\ & \zeta_{i-1} = z_i - \Psi z_{i-1} \end{aligned}\right\} \qquad (6.166)$$

and u_{i-1} and w_{i-1} are independent.

By applying the results of basic estimation theory (as used in Section 6.3.1), formal "filtering", "prediction" and "smoothing" solutions may be derived [21]. However, Bryson and Hendrikson point out a distinction between these formal solutions and the actual solutions they obtain. Since ζ_{i-1} is based on z_i, the prediction of x_i based on ζ_{i-1}, in fact, becomes the best estimate of x_i based on z_i. Consequently, the formal filtering and prediction solutions actually correspond to their single stage smoothing and filtering solutions, respectively. In order to distinguish between formal and actual estimates they adopt the following notation;

$x_{i/k}$ = optimal estimate of x_i given measurement up to and including z_k.

$P_{i/k}$ = covariance of $\hat{x}_{i/k} = E[(x_i - \hat{x}_{i/k})(x_i - \hat{x}_{i/k})^T]$

actual	formal
$\hat{x}_{i/i}$	$\hat{x}_i$
$\hat{x}_{i-1/i}$	$\hat{x}_{i-1}$

where $\hat{x}_i$ and $\hat{x}_{i-1}$ are the formal prediction and estimate based on ζ_{i-1}.

Only the filtering solution is considered here (and used in Chapter 7) and this is summarised as follows;

$$\hat{x}_{i/i} = \Phi\hat{x}_{i-1/i-1} + [D + (\Phi - DH^{r})K_{i-1}](\zeta_{i-1} - H^{r}\hat{x}_{i-1/i-1})$$

where $D = SR^{-1}$, $R = \bar{Q} + HQH^{T}$, $S = QH^{T}$ (6.167)

$$H^{r} = H\Phi - \Psi, \quad \zeta_{i-1} = z_{i} - \Psi z_{i-1}$$

$$K_{i-1} = M_{i-1}H^{rT}(H^{r}M_{i-1}H^{rT} + R)^{-1}$$

$$P_{i-1} = (I - K_{i-1}H^{r})M_{i-1}(I - K_{i-1}H^{r})^{T} + K_{i-1}R\,K_{i-1}^{T}$$

$$M_{i} = (\Phi - DH^{r})P_{i-1}(\Phi - DH^{r})^{T} + Q - DRD^{T}$$

$$P_{i/i} = M_{i}, \quad P_{i-1/i} = P_{i-1}$$

At least two successive measurements in the sequence $\{z_i\}$ must be available before the first estimate can be generated and, before the algorithm can be used at all, some estimate of the correlation parameter Ψ is required.

6.3.5 Simulating Lidar Data

Lidar data in the form of normalised return powers, or the ratio of such quantities, may be regarded as a composite process with, essentially, two components;

(1) The signal itself, which is representative of the dynamic, absorptive state of the atmosphere, integrated over a two-way path length between lidar site and topographic target.

(2) Measurement noise, which corrupts the above signal, and which will have various sources of origin between the photo-detection stage and normalisation procedure.

For a pulsed lidar system, the measurement consists of a sequence of

temporally discrete samples of this composite process. These samples may be identified with the measurement sequences, $\{z_k\}$, discussed above, which form the inputs to the estimation algorithms.

Speckle is regarded as the primary noise source and is multiplicative in nature rather than additive. In Chapter 5, expression (5.7) was used to relate the normalised return power, $P_R(k)$ to the product of the signal component $\bar{P}_R(k)$, and the speckle component $S(k)$. Both $\bar{P}_R(k)$ and $S(k)$ are discrete processes. Therefore, assuming that additive measurement noise terms are negligible compared to speckle, expression (5.7) can be used as a model for generating simulated lidar sequences representative of either normalised power or power ratio data. If the simulated data is to represent log transformed quantities, however, a simple additive relationship can be used instead.

A Fortran program, SIM9.FOR, was written to generate simulated data files and provided a series of options for specifying both the signal and the measurement noise, and also the way in which they were combined. Various signal processes can be derived from the time series models of Section 6.3.2. The principal options provided in SIM9.FOR were:

(1) A constant value

(2) A first order autoregressive model

(3) A first order moving average model

(4) Two serially connected autoregressive processes

(5) Two parallel autoregressive processes.

Figure 6.6 illustrates, in block diagram form, the system models provided by options (2), (4) and (5). These are assumed here to be capable of generating more realistic approximations to the absorptive state of the atmosphere than either a constant value or the moving average

model. Each of the autoregressive components used in the last two examples are first order processes. In principle, model complexity may be increased either by taking a given time series to higher order or else by using lower order models as building blocks. Figure 6.6(a) illustrates both the discrete and the continuous forms for the first order autoregressive process. The continuous representation reveals that an AR(1) model is, in fact, equivalent to an integrating unit provided with feedback. In the absence of feedback, this model reduces to the simple random walk.

Second order processes can be generated by feeding the output of one AR(1) unit into the input of another, but their use as system models means that, in general, estimates must be provided for both autoregressive coefficients. One common solution to this problem is to assume that both coefficients are zero, which yields a process known as the "integrated random walk". Another solution is to set just one of the coefficients to zero, then the model reduces to another common form known as the "smoothed random walk". Such models are also capable of providing non-stationary time series which may be used where past information on the rate of change of x is to be used for a priori prediction [11].

Figure 6.6(c) represents an alternative approach to model generation which may be appropriate where the signal process is expected to possess a complex frequency power spectrum. For example, one of the autoregressive components could be used for generating a random walk (i.e. $\phi_1 = 0$) to provide the trend component, whilst the other contributes an exponentially correlated random variable ($0 < \phi_2 < 1$). Again, however, if used as a system model in the filtering algorithms, estimates may be required for both ϕ_1 and ϕ_2. Also, in general, the state variable for both 6.6(b) and 6.6(c) will be of the vector form rather than a

scalar, thus necessitating more complex implementations of the filtering algorithms.

Time series, representative of speckle measurement noise, may take two forms depending on whether the composite signal is to be generated using a product or a sum. The former will correspond to normalised power data and the latter to the logarithm of the normalised power data. In both cases the speckle sequence will be statistically independent of the signal but, to prevent bias, the first must have an expectation value of unity and the second, an expectation value of zero. In Chapter 5 the probability density function of speckle noise was identified as being closely approximated by the Chi-square function with the parameter, m, equal to 3 or 4 (see expression (5.19)).

Various algorithms have been recommended for generating random variables with a Chi-square distribution [22, 23, and 24] and the version used in SIM9 is attributable to J. Von Neuman. Since the Chi-square sequence of order m has a mean of 2m and a variance of 4m, a speckle sequence based on this series must be modified to provide the required mean, and may also be further modified to match the speckle variances observed in real data. If $\{S_k\}$ is the sequence generated by the Chi-square algorithm then the basic speckle sequence is obtained by transforming each S_k using a relationship of the form,

$$S'_k = a(S_k - 2m) \tag{6.168}$$

which has a zero mean and a variance given by

$$\sigma^2_{S'} = a^2(4m)^2 \tag{6.169}$$

where a is a variable used to control the variance. In this form it may be added to one of the signal processes discussed above to produce an

additive composite signal. Alternatively the mean of the series $\{S_k\}$ may be shifted to unity by adding 1.0 to every S_k and the resulting sequence multiplied by the signal process, using (5.19) to yield a multiplicative composite signal.

If it is further required that the speckle noise sequence be sequentially correlated, the simplest method of providing this is to substitute (6.168) for the noise term, a_t, in the autoregressive process of equation (6.102) to obtain the correlated series

$$S''_k = \phi S''_{k-1} + S'_{k-1} \tag{6.170}$$

This will have an autocorrelation function described by (6.116),

$$\rho_k = \phi^k \quad \text{for } k > 0$$

and a variance given by (6.110),

$$\sigma^2_{S''} = \sigma^2_{S'}/(1 - \phi^2) = a^2(4m)^2/(1 - \phi^2) \tag{6.171}$$

As before, the speckle sequence (6.170) may be additively combined with a signal process or, by shifting the mean of this series to unity, each term can be used to form a product according to expression (5.19).

Since the linear Kalman filter requires a linear measurement model, the simulations considered in Chapter 7 use additive composite signals to represent log transformed data directly. The alternative is to generate a multiplicative simulation and log transform this, but since the transformation is non-linear, the advantages of using time series with pre-defined statistical properties would be lost.

Simulated data files in which the component processes are completely specified find immediate application in the testing of estimation algorithms since the estimate is readily compared with the true signal.

This is particularly true for the Kalman filter where the noise covariance matrices, Q and R, will be known in advance. Checks can be made on the filtering algorithms to ensure that the steady state values of the error covariance and Kalman gain matrices (P_∞ and K_∞) converge to their predicted values. Having established that the algorithm coding is error free, estimates of Q and R, may be obtained using the Mehra techniques for instance and compared with the known values. Steady state values of P_∞ and K_∞ based on the estimates, $\hat{Q}$ and $\hat{R}$, may then be compared with the values obtained using Q and R. Similar procedures may be employed with the Bryson and Hendrikson algorithm.

It is apparent, in fact, that there are many ways of using simulation to determine the sensitivity of an estimation algorithm to variations in any of the model parameters. If differences in model complexity are also introduced, between simulation and filter, the potential area for investigation rapidly expands. Chapter 7 includes a brief evaluation of filter performances based on four simulated data files. Estimated values of Q, R, P_∞ and K_∞ are compared with their anticipated values, using the basic linear (discrete) Kalman filter, the Mehra algorithm, and the Bryson and Hendrikson filter for both white and correlated measurement noise sequences.

REFERENCES

[1] MENYUK, N., KILLINGER, D.K. and MENYUK, C.R. "Limitations of signal averaging due to temporal correlation in laser remote-sensing measurements", Applied Optics, 21(18), Sept. 1982, pp. 3377-3383.

[2] GELB, A. (Ed.) "Applied Optimal Estimation", The MIT Press, ISBN 0 262 70008 5, 1974.

[3] ZRNIC, D.S. "Mean power estimation with a recursive filter", IEEE Transactions on Aerospace and Electronic Systems, AE5-13(3), May 1977, pp. 281-289.

[4] PIPES, L.A. and HARVILL, L.R. "Applied Mathematics for Engineers and Physicists", McGraw-Hill, 1970.

[5] BOZIC, S.M. "Digital and Kalman Filtering", Edward Arnold (Publishers) Ltd., 1979, pp. 92-94.

[6] SWERLING, P. "First order error propagation in a stagewise smoothing procedure for satellite observations", J. Astronautical Sciences, 6, 1959, pp. 46-52.

[7] KALMAN, R.E. "A new approach to linear filtering and prediction problems", J. Basic Engineering, 82D, March 1960, pp. 35-46.

[8] KALMAN, R.E. and BUCY, R.S. "New results in linear filtering and prediction theory", J. Basic Engineering, March 1961, pp. 95-108.

[9] BLUM, M. "A stagewise parameter estimation procedure for correlated data", Numer. Math., 3, 1961, pp. 202-208.

[10] BATTIN, R.H. "A statistical optimising navigation procedure for space flight", ARS Journal, 32, 1962, pp. 1681-1696.

[11] YOUNG, R. "Recursive Estimation and Time Series Analysis - An Introduction", Springer-Verlag, 1984.

[12] CADZOW, J.A. "Discrete Time Systems: An Introduction with Interdisciplinary Applications", Prentice-Hall, 1973.

[13] BENEDICT, J.R. and BORDNER, G.W. "Synthesis of an optimal set of radar track-white-scan smoothing equation", IRE Transactions on Automatic Control, , July 1962, pp. 27-32.

[14] RYE, B.J. Private communication.

[15] BOX, G.E.P. and JENKINS, G.M. "Time Series Analysis: Forecasting and Control", San Francisco: Holden-Day, Inc., 1970.

[16] VANDAELE, W. "Applied Time Series and Box-Jenkins Models", Academic Press, ISBN 0 12 712650 3, 1983.

[17] MEHRA, R.K. "On the identification of variances and adaptive Kalman filtering", IEEE Transactions on Automatic Control, AC-15(2), April 1970.

[18] BÉLANGER, P.R. "Estimation of noise covariance matrices for a linear time-varying stochastic process", Automatica, 10, 1974, pp. 267-275.

[19] CAREW, B. and BÉLANGER, P.R. "Identification of optimum filter steady-state gain for systems with unknown noise covariances" IEEE Transactions on Automatic Control, AC-18(6), December 1973, pp. 582-587.

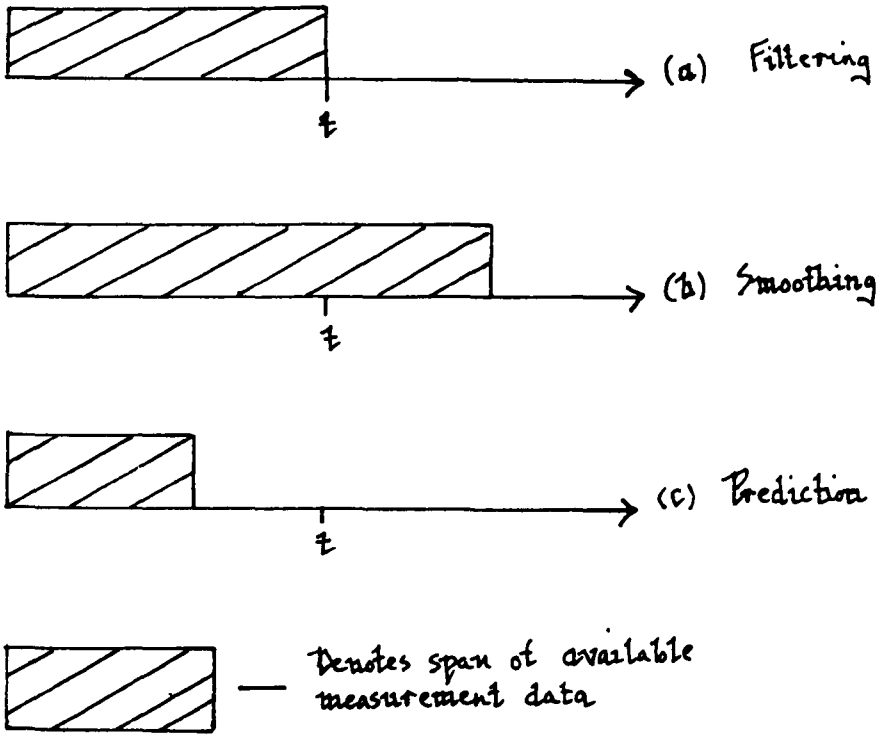
[20] NEETHING, C. and YOUNG, P. "Comments on "Identification of optimum filter steady-state gain for systems with unknown noise covariances"", IEEE Transactions on Automatic Control, , October 1974, pp. 623-625.

[21] BRYSON, Jr. A.E. and HENDRIKSON, L.J. "Estimation using sampled data containing sequentially correlated noise", J. Spacecraft, 5(6), June 1968, pp. 662-665.

[22] VON NEUMAN, J. "Various techniques used in connection with random digits", U.S. NBS Appl. Math. Ser., AMS12, 1951, pp. 36-38.

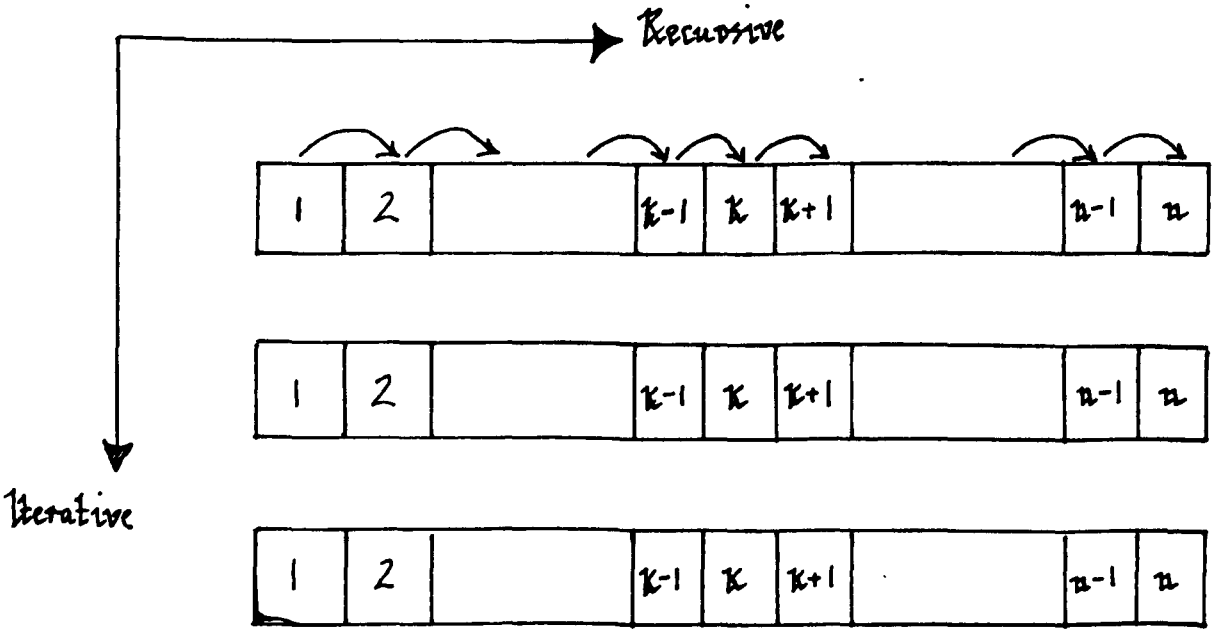
[23] KAHN, H. "Applications of Monte Carlo", Rand Corp. Rep. RM-1237-AEC, April 1956.

[24] MARSAGLIA, G. Ann Math. Stat. 32, 1961, pp. 899-900.



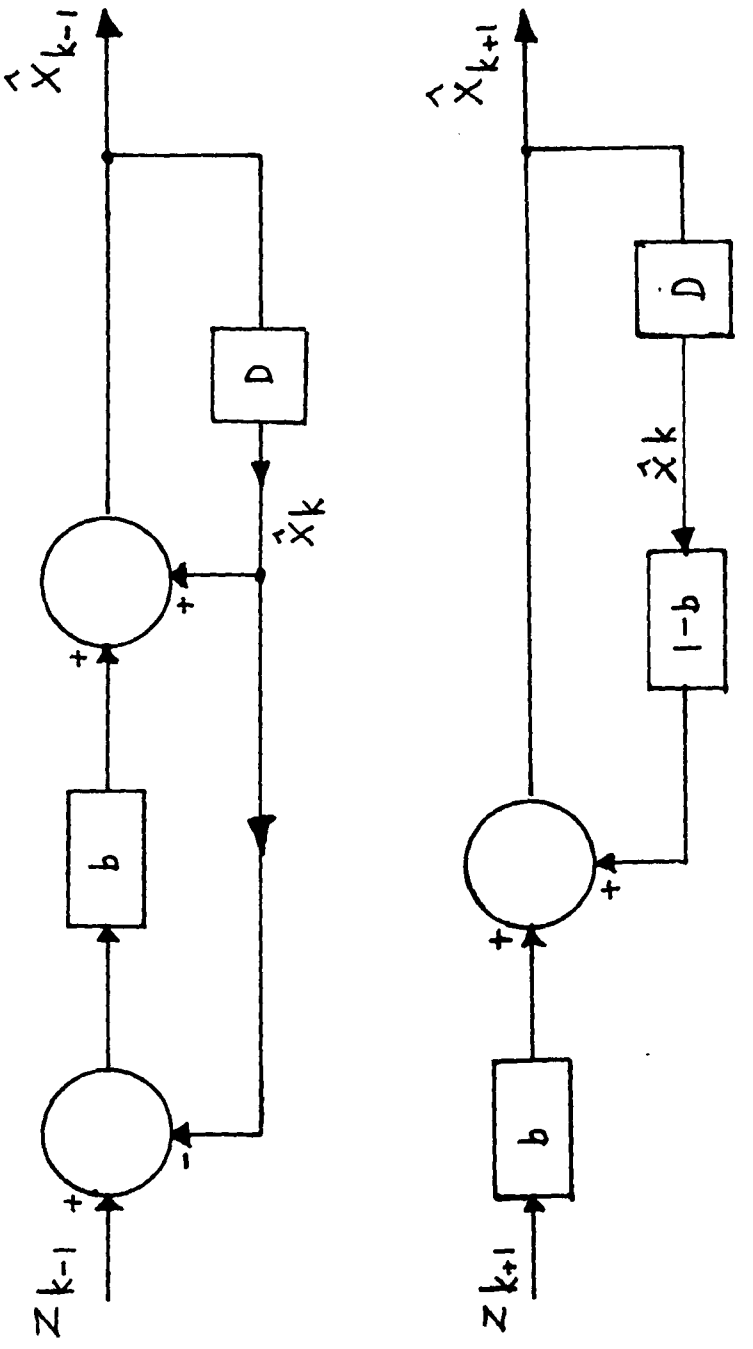
Definition of Filtering, Smoothing and Prediction

Fig. 6.1



Recursive and Iterative Data Processing Schemes

Fig. 6.2



Representations of the First Order Recursive Filter

Fig. 6.3

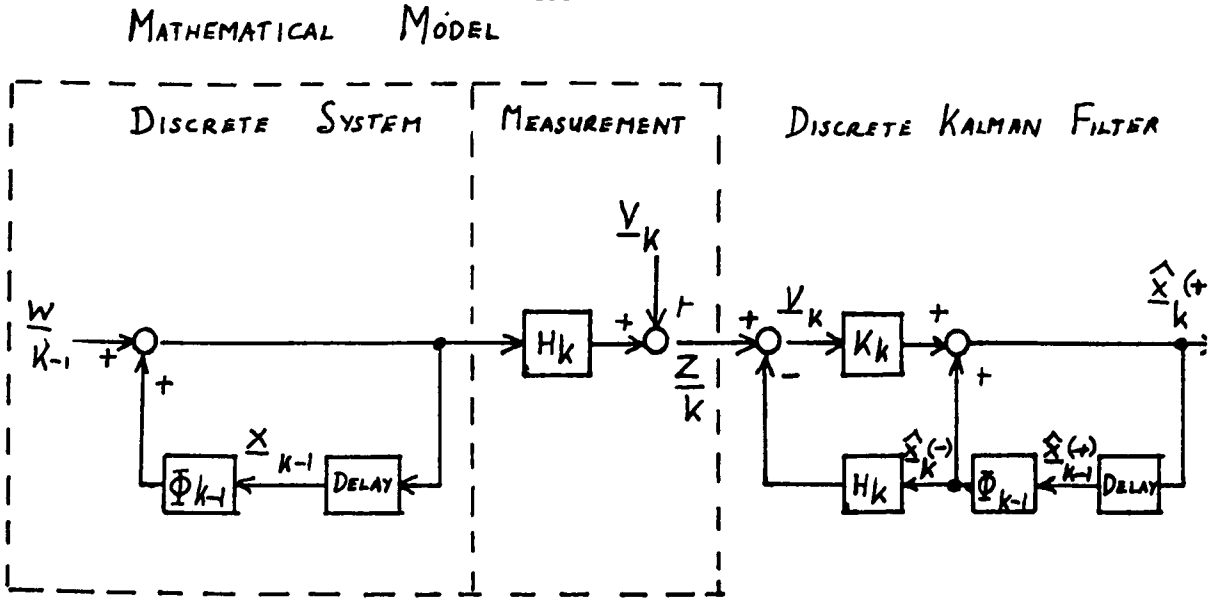


Fig. 6.4 System Model and Discrete Kalman Filter (after Gelb, [ref 2]).

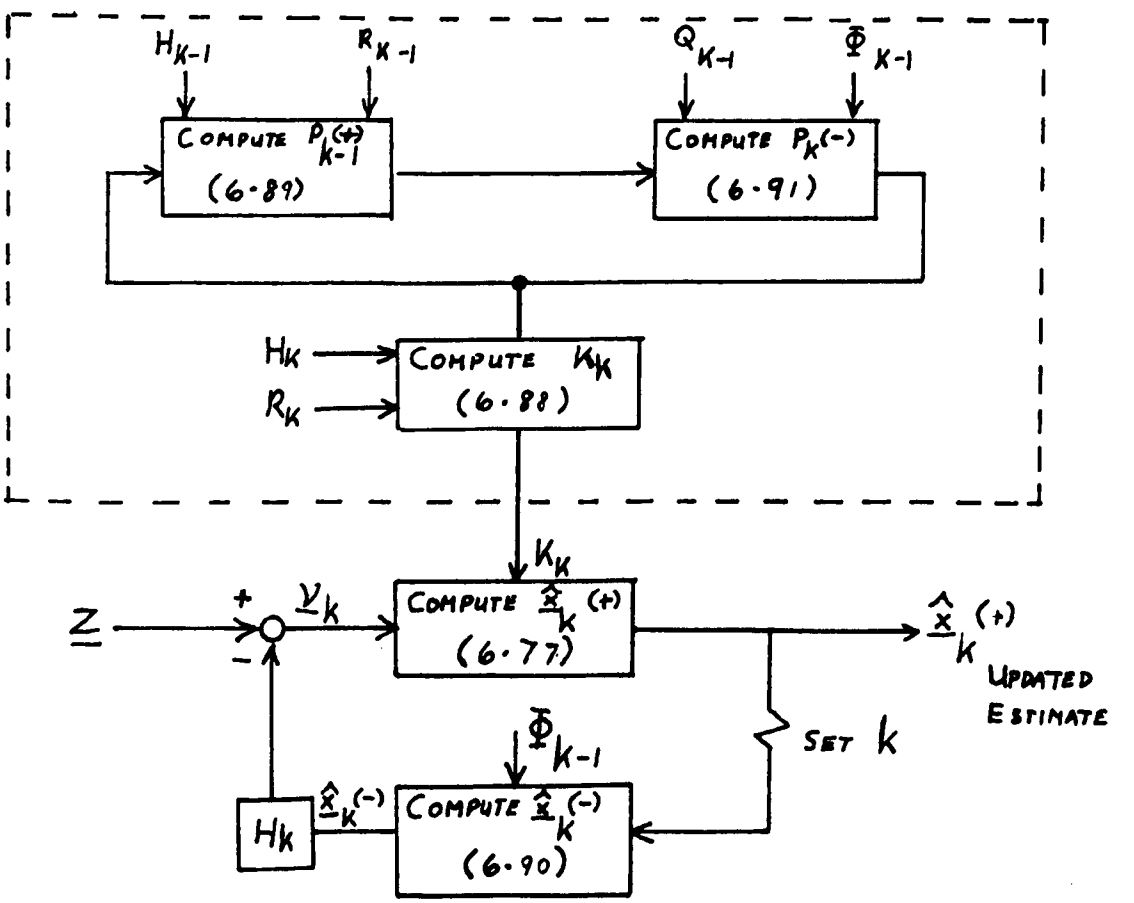
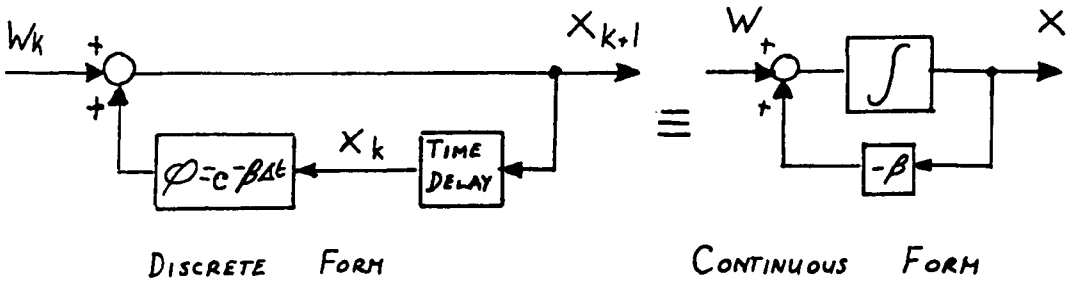
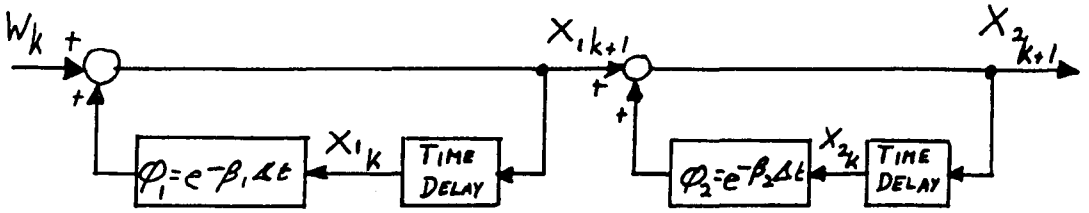


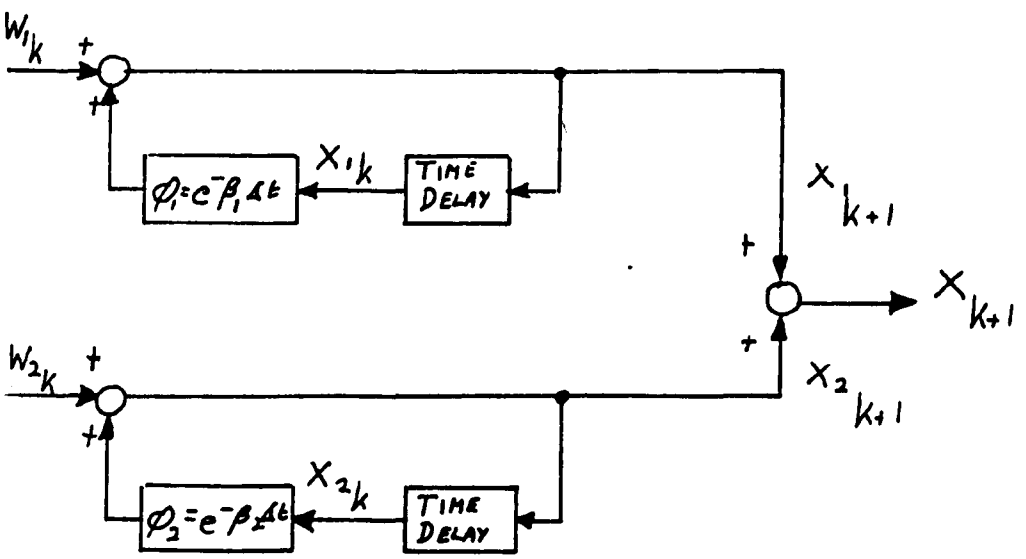
Fig. 6.5 Discrete Kalman Filter Information Flow Diagram (after Gelb, [ref 2])



(a) AR(1) Model



(b) AR(1) Model



(c) Parallel AR(1) Model

Fig. 6.6

CHAPTER 7.0

DATA ANALYSIS

Prior to any data analysis of multiple lidar returns, it was difficult to predict exactly how many pulse pair returns would be necessary for the subsequent application of various estimation methods. In fact, as described earlier in Chapter 6 and evidenced later in this chapter, the successful application of any estimation method tends to be an iterative process in which the most elementary methods are applied first in order to gain information which need not always be of a quantitative nature.

Typical normalised variances are dependent on the amount of aperture averaging affected by the optical receiver configuration of a particular lidar system but, in the heterodyne detection limit, they reach a maximum of unity (approximately) [1]. For this worst case variance, and assuming a constant signal, precisions of the order of 1% can be achieved by averaging 10,000 pulse pairs, since variance reduction for a constant signal is inversely proportional to the root of the number of samples used.

A central assumption in this work, however, is that the signals to be estimated are not constant over the duration of the experiment and that simple averaging methods are not valid. Support for this assumption was already available in the work by Killinger [4] in which he demonstrates the departure from the $N^{-\frac{1}{2}}$ dependence using databases consisting of the order of 20,000 pulse pairs. His method of analysis will be discussed later but database populations of this magnitude appear appropriate for his methods.

Clearly, however, if traditional estimation methods are to be applied in which averaging techniques are implicit, large sample quantities are necessary. This will be true regardless of whether a single value estimate, or a time dependent estimate is to be extracted from the measurement sequence. Optimal estimation methods will help minimise the number of samples necessary to achieve a desired accuracy but sample records beyond this minimum will, in general, still be desirable for the analysis of the temporal evolution of the estimated parameter.

The above considerations indicated that each measurement sequence consists of at least a few thousand shots. An upper limit of 10,000 shots was, however, imposed by constraints of a practical nature. Although the dual laser system was capable of high rep rates for short durations, an optimum prf of approximately 10 Hz was used since this tended to maximise the number of firings before arcing occurred. Also, beyond a measurement duration of $\sim$ 25 minutes, thermal instabilities tended to necessitate re-alignment and the cryogenic cooling of the LTT detector sometimes caused misting of its zinc selenide window. Therefore, to maximise the number of valid measurements, a standard measurement of 10,000 samples was decided upon which required approximately 17 minutes to collect.

Preparation of the lidar system prior to each experiment required a considerably longer period of time than the measurement duration. Between measurements it was necessary to:

(i) perform some preliminary analysis on the data of the previous measurement;

(ii) evacuate both low and high pressure gain sections;

(iii) fill them with the appropriate laser gas mixture and check for correct lasing action in all four gain sections;

(iv) line tune each hybrid to the wavelength selected for the next experiment;

(v) align both laser outputs into the transceiver; and finally

(vi) check the correct operation of both signal processing electronics and the data capture software. Such a procedure tended to limit the number of measurements to a maximum of three per day but with an average nearer two per day.

In order to ensure that no unrecorded modifications were made to the system between measurements a formal method of recording all system variables and ambient conditions was devised. Temperature, humidity, date, time, amplifier gains, laser lines selected, voltages used, output energies, waveform recorder sample interval, intra pulse pair delay, and data file names are typical examples of the information recorded.

7.1 Selected Measurements

The sequence of measurements performed can be partitioned into two phases. The first was essentially a fine tuning phase in which many measurements were made with each laser tuned to the same wavelength. Primary goals were the refinement of the transmitted pulse normalisation hardware and adjustment of the dynamic range to levels comparable with the return pulses, optimisation of the signal processing electronics, and refinement of the data capture software.

Phase two formed the experimental period, proper, and three categories of experiment were planned:

(i) Having refined the system as far as was possible within practical time constraints, an immediate requirement was to test the system

with both lasers tuned to the same wavelength. Under ideal conditions, of course, both channels should yield identical information about the absorption in each channel, but differences of systematic origin had been anticipated. Measurements of this sort would help quantify the magnitude of these errors, but would also be useful in assessing the estimation techniques applied later in this chapter since their performances on the data from each channel could be expected to be highly correlated. The wavelength selected for these measurements was the R18 transition at 10.2604 microns. This was chosen primarily because of its close proximity to other wavelengths (later to be used for absorption experiments) which tended to minimise the amount of laser/transceiver telescope alignment necessary after each re-tune. Other influential factors, however, were its relatively high line strength and relatively low absorption by H_2O.

(ii) A principal target for any CO_2 lidar system intended for gas concentration measurements is H_2O (if within the tunable range of the lasers used), simply because of its abundance. It is also often the principal interfering species when measuring the concentration of other species. Laser lines selected for the monitoring of H_2O were extracted from various authors' recommendations, together with their published values for the absorption coefficients [Table 1.1, References 2,3]. The R20 transition at 10.2466 microns is usually recommended as the measurement wavelength, but either R18 at 10.2604 microns or R16 at 10.2744 microns appear to be reasonable choices for the reference wavelength. Measurements were made using all three wavelengths and the results based on these files will be presented in this chapter.

(iii) Of particular importance as atmospheric pollutants, however, are ammonia (NH_3) and ethylene (C_2H_4), for reasons discussed in Chapter 1. Measurements of these gases were attempted, again using recommended lines and quoted absorption coefficients. For ammonia, the recommended laser transitions for the measurement at CO_2 wavelengths is often R8 at 10.3337 microns with R12 at 10.3035 microns as the reference. Ethylene monitoring requires the use of P branch transitions with P14 at 10.5321 microns as the measurement wavelength and P28 at 10.6746 as the reference wavelength. Results based on these measurements will also be presented in this chapter.

Before proceeding, first to the application of sub-optimal methods to the above measurement data, it is worth stressing here that the achievement of high precision concentration measurements is not a primary objective of this work. Uncertainties in the differential absorption coefficient due to species interference, the effect of pressure and temperature variations on these absorption coefficients (known or otherwise), and systematic errors were discussed in Chapter 5 and tend to preclude accuracies better than a few per cent.

Of fundamental importance, however, is the difficulty in corroborating gas concentration estimates based on an atmospheric path length of 2 x 1.8 km. A conventional 'hair based' hygrometer was used to monitor humidity at the lidar site and could therefore be used to compare relative humidity estimates resulting from the measurements. Discrepancies between these estimates and the hygrometer reading are difficult to interpret, however, since the latter instrument is limited in accuracy to a few per cent and the comparison would be between a point sensor and a path integrated measurement in close proximity to the open water of the Humber

estuary. For the ammonia and ethylene measurements, no means of corroboration were available but ambient levels of both gases could have been greater than the global average due to the close proximity of various industries.

Although there is a clear limit to the precision with which any absolute concentration measurement can be stated under these circumstances it is still an important independent requirement to be able to estimate accurately normalised return powers, their ratio and of a more direct interest, the natural logarithms of both quantities. Where appropriate, therefore, estimations of these quantities are translated into, and presented as, concentration profiles (and relative humidity profile for H_2O) but their primary significance is as an indication of the temporal evolution of the absorptive state of the atmosphere over the duration of an experiment.

Many of the graphs included in this chapter contain information in their titles relating to the version number of the program used to process the data, processing parameters used, and data file names. These have been preserved here to provide a reference should any subsequent use be made of these files.

7.2 Preliminary Analysis

7.2.1 The Estimation Problem

Menyuk and Killinger appear to be the first to explicitly state a problem fundamental to lidar estimation, and in fact relevant to any other situation in which a parameter must be estimated from noisy data by taking many samples over a period of time during which that parameter may be expected to vary [4]. They used large measurement sequences,

in excess of 20,000 normalised returns, to demonstrate the departure from the $N^{-\frac{1}{2}}$ dependence often assumed to exist for the standard deviation of an estimate based on an average of N samples. As indicated when establishing the relationship (6.6), this will only be true in certain cases (if, for example, the estimated parameter remains constant whilst the samples are being collected).

Figure 7.1 is a confirmation of their findings based on a measurement sequence of 10,000 single channel normalised lidar returns. Their method of analysis was repeated for this measurement and the figure clearly illustrates how the standard deviation of the estimate shows significantly less improvement as the number of samples increases. Logarithmic transformations (base 10) of both axes have been applied to conform to the method of presentation adopted by Menyuk and Killinger. The curve, (b), was derived by applying the segmental averaging techniques of Section 6.1.2 in which the total number of samples, N, is divided into M (= N/k) subsets of data. Each k sample forms M independent averages and the standard deviation, σ_K, is calculated using

$$\sigma_k^2 = \frac{1}{M}\left[\left(\frac{I_1+I_2+\quad+I_k}{k}\right)^2 + \left(\frac{I_{k+1}+\quad+I_{2k}}{k}\right)^2 + \quad + \left(\frac{I_{N-k+1}+\quad+I_N}{k}\right)^2\right] \tag{7.1}$$

where $I_i = \dfrac{P_i - \bar{P}}{\bar{P}}$

and $\bar{P}$ is the average formed using the entire data set (N values).

This result is a direct consequence of applying averaging techniques to data sequences having non-zero autocorrelation functions. It indicates

that successive samples are not independent, and that this lack of independence may be due to a non-stationary absorptive state of the atmosphere and/or temporal correlation in the measurement noise. Figure 7.1 is typical of all the lidar data analysed in this way at Hull University and indicates the existence of a positive autocorrelation function. It is worth noting here that, in a hypothetical situation in which the autocorrelation function at all lags were to collapse towards zero, curve (b) would then approach curve (a). It has been suggested that, if the autocorrelation function should become negative, then averaging techniques could actually provide an improvement over the theoretical $N^{-\frac{1}{2}}$ dependence, and this situation is indicated in the hypothetical curve, (c) [5].

Menyuk and Killinger's most important conclusion, however, is that, if the autocorrelation function,

$$\rho_i = \frac{1}{\sigma^2 (N-j)} \sum_{k=1}^{N-j} I_k \, I_{k+j} \qquad (7.2)$$

where $$\sigma^2 = \frac{1}{N} \sum_{k=1}^{N} I_k^2$$

is any monotonically decreasing, or constant function of delay time, then the variance reduction achievable by averaging reaches the fundamental limit,

$$\sigma_k^2 \geqslant \sigma^2 \rho_{k-1} \qquad (7.3)$$

Expressed as a variance reduction ratio, this limit becomes

$$K = \frac{\sigma_k^2}{\sigma^2} = \rho_{k-1} \qquad (7.4)$$

Therefore, "as long as the monotonically decreasing temporal correlation coefficient, ρ_{k-1}, has a finite value for a given time interval (lag), that value limits the improvement in the standard deviations that can be obtained by signal averaging regardless of the number of pulses averaged during that interval" [4].

It will be seen later that some of the measurements analysed exhibit finite autocorrelation functions as large as $\rho_{k>10} \approx 0.1$ which would limit averaging techniques to a maximum useful sample of $k \lesssim 10$, yielding an accuracy no better than approximately 30%!

7.2.2 Sub-optimal Estimation Techniques

The sub-optimal methods applied in this chapter are restricted to just two techniques: the $\alpha\beta$ tracker and the running average. Both can be made to produce very similar results and leave only one parameter for adjustment according to the user's discretion. Restricting the techniques applied here has the advantage of avoiding multiple plots of estimates which differ only slightly in consequence as well as appearance.

The $\alpha\beta$ tracker has already been used in Chapters 2 and 5 as a means of filtering data to reveal salient features of an estimated process. Examples presented in Figures 5.4(b) and (c) illustrated its application to the filtering of raw digital values corresponding to the transmitted reference and return powers, whilst Figure 5.4(d) is an example of its application to the normalised power values formed by ratioing these digital values prior to filtering. This filtering procedure was in fact repeated immediately after each measurement to provide an initial check on the data and to reveal the presence or absence of any trend.

Figures 7.2(a), (b), (g) and (h) are four representations of the state of relatively raw, unfiltered data for a particular measurement

using the R20/R18 lines. The first two illustrate the normalised powers measured in each lidar channel, here denoted 'A' and 'B', but (g) and (h) respectively show their ratio and the logarithm of the ratio. Each of these plots illustrates 1000 points extracted by a graphics routine which samples every tenth value from the original data set.

Figures 7.2(a) and (b) use identically scaled axes to permit direct comparison of graphs presented in sequential order, and this method of presentation will be adopted wherever appropriate throughout Chapter 7. Ordinate axes expressing normalised power, or functions thereof, refer to the dimensionless quantities formed by ratioing the digitised voltage corresponding to the received and transmitted powers per channel (details of the normalisation method are presented in Chapter 2 and further discussed in Chapter 5). All graphs forming Figure 7.2 are derived using the same measurement file (SIFT5.045).

A significant difference in the "mean" signal level is immediately apparent between the two channels (7.2(a) and (b)) and this is directly attributable to the fact that this measurement was conducted using a pair of lines suitable for water vapour measurement; the R20 line corresponds to channel B and has an H_2O absorption coefficient nearly an order of magnitude greater than that of the reference line, R18 corresponding to channel A (see Table 7.1). Figures 7.2(a) and (b) also clearly demonstrate the multiplicative nature of speckle noise. When forming the ratio of the normalised returns in each channel, the resulting quantity can therefore be expected to exhibit a large dynamic range and this is, in fact, illustrated in Figure 7.2(g). Forming the log-ratio then compresses the dynamic range as shown in Figure 7.2(h).

By comparison, Figures 7.3(a) and (b) show the normalised power returns obtained using both lasers tuned to the same line (R18) for the data file, SIFT5.026. Correlation coefficients obtained for data in this state (that is, after normalisation but prior to filtering) are typically close to 0.9 and, by superimposing both figures for the particular example illustrated, it is possible to detect this correlation visually.

7.2.2.1 Estimates of Normalised Power Data

Figures 7.2(c) and (d) present the results of applying respectively, an $\alpha\beta$ tracker (critically damped - see Section 6.2.2) and a running average to the normalised powers recorded in channels A and B corresponding to the water vapour measurement lines R18 and R20. In each case a single parameter is used to control the degree of filtering - alpha for the $\alpha\beta$ tracker and n for the running average (equivalent to the number of samples averaged to form each estimate in the curve - see Section 6.1.3). The values of α and n used in Figures 7.2(c) and (d) were selected heuristically by applying different values to various data files, specifically to channels for which relatively high returns were obtained, and attempting visually to achieve a balance between "over" and "under" filtering. In applying these subjective criteria, the primary goal was to assess the presence or absence of trend.

A single value of $\alpha (= 3.0 \times 10^{-3})$ and n (= 200) has been used for each channel but, because the speckle induced variance is significantly different between channels, the degree of filtering which results also differs considerably. For the purposes of discerning trend on a time-scale shorter than the duration of the experiment, the estimate for

channel A would seem sufficiently responsive, whereas the channel B estimate would appear more suitable for detecting changes in absorption occurring over a duration greater than that of the experiment.

It is possible to select the values of α and n such that the variance reduction is identical for each method. This has not been attempted in Figures 7.2(c) and (d). Using the values of α and n indicated above, running average, in fact, produces a "noisier" estimate than the $\alpha\beta$ tracker. This difference is discernible from the figures, but less obvious is the lag which results from differences in the filter responsivity which, in turn, are dependent on the selected values of α and n. A low value of α, or n, produces a very responsive, but noisy, estimate and vice versa. The $\alpha\beta$ tracker, of course, also provides an estimate of this responsivity since the recursive algorithm (Section 6.2.2) generates both zero and first order differentials of the estimated parameter.

Other values of α and n could have been applied here but, without some quantitative guidance as to whether one value produced a "better" estimate than another, the only result would have been a multitude of graphs illustrating various degrees of filtering. It could be argued that some guidance is available, for instance, via the variance reduction ratio quoted for the $\alpha\beta$ tracker in Section 6.2.2. Figure 7.4 is a plot of this reduction ratio (K) as a function of α. If $\alpha = 3.0 \times 10^{-3}$, the corresponding value of K is approximately 2.0×10^{-3} which, for a normalised power variance of 25% (typical for the Hull lidar system), would appear to offer an improvement in accuracy down to approximately ±2%.

Figure 7.4, therefore, would seem to imply that further improvement can be obtained simply by going to even smaller values of α until, in the limit as $\alpha \to 0$, the estimate is entirely error free. This is true,

however, only for the special case in which the estimated parameter is actually a constant. For any parameter exhibiting time dependency there will be an optimum, finite, non-zero value of α dependent on the frequency-power spectrum of the estimated parameter (or its statistical/stochastic properties, depending on the type of analysis applied).

Application of filtering in excess of the optimum represents a degradation in precision in just the same way as under filtering; in the former case the excess noise manifests itself in an inability of the filter to track sudden changes in the parameter being estimated. Under filtering leaves an excess of noise, originating in the measurement process, remaining in the estimated parameter.

If the $\alpha\beta$ tracker and running average are now applied to the data file for which both lasers were tuned to the same wavelength (SIFT5.026), the results obtained are as presented in Figure 7.3(c) and (d). Both methods used the values of α and n employed previously. Although highly correlated, as expected, both figures reveal a discrepancy between the absolute values of the estimates of up to, approximately, 10%. The exact magnitude of this discrepancy will obviously depend on the estimation method used but the presence of a systematic error is clearly indicated.

Various possible sources of such errors were anticipated in Chapter 5 but the consistency with which the estimate of channel B falls below that of channel A (for both filters) suggests a relatively constant disparity in the percentage energy losses occurring in one or more of the following:

(a) the transmitter optics,

(b) the receiver optics, and/or

(c) the reference beam optics.

The predicted energy losses, based on simple beam waist considerations and presented in Figure 5.1, suggested a discrepancy of between 5% and 10% due to the diameter of the flat used in transmitter telescope.

Some variation in the apparent magnitude of this discrepancy, however, is also evident over the duration of the measurement, perhaps attributable to changes in the beam mode energy distribution superimposed upon a primary cause of disparity of the type indicated above.

7.2.2.2 Estimates of the log of the Normalised Power

Applying the logarithmic transformation (base e) to the data of Figures 7.2(a) and (b), before using the αβ tracker and running average, yields the estimates illustrated in Figures 7.2(e) and (f). Here, again, the selected values of α and n are maintained at the same values as before for reasons of consistency. Perhaps the most salient feature of these two estimates is the extent to which the non-linearity of the log transformation has compensated for the different variances in the original data from channels A and B. This is attributable to the approximation [6],

$$\sigma[\ln(x)] \simeq \frac{\sigma(x)}{\bar{x}} \qquad (7.5)$$

which states that the standard deviation in the log of a variable is approximately equal to the normalised standard deviation of the variable itself. Under this transformation, therefore, the estimates appear to have equivalent variances. In other respects, however, the same comments, applied to the αβ tracker and running average in the previous section, apply here also.

For the dual channel, single wavelength measurements, estimates of the log transformed data are shown in Figures 7.3(e) and (f). Further consideration will be given to this file in Sections 7.2.3 and 7.3.

Figures 7.5(a) through to 7.5(h) illustrate running average estimates of log-transformed, normalised power data for other files, most of which will also be considered later in this chapter. A single value of n (= 200) was used to produce all of the estimates in these figures, with the exception of 7.5(g) for which n = 50. This example may be compared with the previous figure, 7.5(e), which presents the running average estimate (n = 200) for the same data file, SIFT5.044. Table 7.1 lists these file numbers, together with the laser lines used, their absorption and differential absorption coefficients for H_2O, CO_2, NH_3 and C_2H_4, and example values of the log ratios expected for typical concentrations of these gases.

7.2.2.3 Estimates of the log of the the Ratio of the Normalised Powers

Once again, the same values of α and n have been applied to generate the estimates presented in Figures 7.2(i) and (j). In forming the ratio, features characteristic of channel A data now dominate the estimate (c.f. with 7.2(c) and (d)) due to the higher mean signal level detected in this, the reference channel.

Concentrations and relative humidity profiles could, obviously, now be derived from these two estimates. This final transformation of the data will, however, be deferred until some quantitative statement can be made regarding the precision of the estimate on which they are based (Section 7.3). Before proceeding to the application of techniques which will help identify characteristics of the signal relevant to the task of

building an optimal filter, the αβ tracker and running average are applied in a final qualitative evaluation, this time using simulated data.

7.2.2.4 Simulated Data

Techniques used in simulating data were discussed in Chapter 6 but, in essence, they assume that two dominant processes contribute to the composite signal detected and recorded as lidar data. The first is the variation in the power, or power ratio, attributable purely to random changes in atmospheric absorption, and the second is measurement noise which is produced when these pulsed powers are optically averaged and electronically detected. The latter is dominated by the phenomenon of speckle and is due to the partially coherent field of laser radiation collected by the receiver telescope which, when imaged at the active element of the detector, produces the time dependent granular pattern of illumination known as speckle. Although the mean power associated with this optical field may be constant, pulse to pulse variations in the power detected are significant due to changes in the imaged speckle pattern between shots.

Figures 7.6(a) and (b) illustrate two simulated sequences of log power data, each of which assumes a constant value of -1.5 for the log of the normalised power in the absence of speckle. Both examples display an additive speckle sequence of constant normalised variance, generated using the Chi-square random variable with m = 3 (see Section 6.3.5). However, whereas in Figure 7.6(a) this sequence is temporally uncorrelated ("white" speckle), in 7.6(b) the same sequence was processed to exhibit exponential correlation ("coloured" speckle) and has a correlation coefficient which decays to 0.1 after 5 "shots".

If the $\alpha\beta$ tracker is now applied to both sets of data, using the same value of $\alpha (= 3.0 \times 10^{-2})$ as used before, the results obtained are as illustrated in Figures 7.6(c) and (d). Each of the samples forming the correlated speckle example are not completely independent, with the result that the effectiveness of any estimation technique applied to such data is actually diminished compared to an equivalent signal for which the speckle is uncorrelated. This difference appears in these two figures as discrepancy in the visually discernible variance of the estimates; the correlated speckle case exhibits a higher variance than the alternative with uncorrelated speckle.

Figures 7.6(e) and (f) reproduce these results using the running average with n = 200. Since the running average estimator, with n = 200 has a higher responsivity but lower variance reducing properties than the $\alpha\beta$ tracker with $\alpha = 3.0 \times 10^{-3}$, the differences in variance reduction achieved is even more conspicuous in these two examples.

A constant mean was used in the previous example to provide an uncomplicated demonstration of the significance of correlated measurement noise in limiting the number of samples which can be regarded as effectively independent. It is, however, an unrealistic model of the real absorption signal. During a prolonged measurement, the absorption characteristics of the atmosphere, at any CO_2 laser wavelength, over a two way path length of 3.6 km, are unlikely to remain constant in the presence of an absorbing species such as H_2O which has a non-uniform temporal and spatial distribution. In general, it can be expected to vary in some non-deterministic manner with statistical properties dependent on ambient conditions such as wind, turbulence and advection, manifesting changes over periods both long (equivalent to trend) and short compared with the duration of the measurement.

Methods for quantitatively identifying the statistical properties of discrete measurements of the absorptive state of the atmosphere were discussed in Chapter 6 and will be applied later in this chapter. However, since complex signal model identification is not a primary objective of this work, a simple non-deterministic model is employed here which will be justified later and which is capable of providing an approximation to the randomly varying absorption characteristics of the atmosphere. This model is the "random walk" which, in discrete notation has the form

$$x_{k+1} = x_k + w_k \tag{7.6}$$

where w is a zero mean gaussian noise sequence having variance, q. The random walk has a power spectral density function which is uniform out to some maximum cut off frequency governed by the driving noise variance, q. It also forms a non-stationary time series, having a variance which increases as k_q and is therefore infinite for an infinite series. For a finite sequence, however, the non-stationarity of the random walk is useful for modelling trend.

Figure 7.7(a) illustrates a random walk signal initiated at x_o = -1.5 and having a driving noise variance of $q = 10^{-5}$. This value of q is somewhat arbitrary but produces a plausible representation of the way in which the logarithm of the return power (in the absence of speckle) may be expected to vary. The speckle sequences used in Figure 7.6(a) and 7.6(b) for the constant signal case have been added to this random walk to produce the composite signals shown in Figures 7.7(b) and 7.7(c) (7.7(c) is the correlated version).

The $\alpha\beta$ tracker estimates ($\alpha = 3.0 \times 10^{-3}$) for both speckle versions are presented in Figures 7.7(d) and 7.7(e) where, once again, the larger

variance of the correlated speckle estimate, (e), is discernable. As before, the running average estimates (n = 200) are also presented here for both types of speckle (Figures 7.7(g) and (h)) and, in common with Figures 7.6(e) and (f), differences in the estimated variances between correlated (7.7(h)) and uncorrelated (7.7(g)) speckle are more pronounced. The speckle independent random walk signal of Figure 7.7(b) has been included in all figures, 7.7(d) to 7.7(h) (broken line) to facilitate these comparisons.

Both of these simulated data files, for the random walk log power signal with correlated and uncorrelated speckle, will be used again in Section 7.3 when evaluating the optimal estimation method.

7.2.3 Model Identification

In applying the optimal estimation techniques of Kalman filter theory, a central requirement is to be able to specify the signal and measurement process according to the formal methods discussed in Section 6.3.1. The discrete formulation of the signal process, in a generalised form, was presented there as

$$\underline{x}_{k+1} = \Phi_k \underline{x}_k + \Gamma_k \underline{w}_k + \Lambda_k \underline{u}_k \tag{7.8}$$

and for the measurement,

$$\underline{z}_k = H_k \underline{x}_k + \underline{v}_k \tag{7.9}$$

where x denotes the state vector, Φ is the state transition matrix, $\underline{w}$ and $\underline{v}$ are vectors of random processes, Γ is the noise covariance matrix, $\underline{u}$ is a vector of deterministic (control) inputs, Λ, its associated matrix, and H is the observation matrix.

For the scalar, single channel processing requirements discussed here, the above relationships reduce to the scalar form,

$$x_{k+1} = \phi_k x_k + g_k w_k \tag{7.10}$$

$$z_k = h_k x_k + v_k \tag{7.11}$$

where the vector/matrix notation has been dropped because it is no longer required, and where the deterministic control input term has been omitted because it does not apply. In pursuing the Box-Jenkins approach, time series formulations of the Auto Regressive, Moving Average (ARMA) form were considered adequate models for approximating the random process, assumed to be governing the temporal dependence of the absorptive properties of the atmosphere. The general form of the ARMA (p,q) model was represented by

$$x_k = \phi_1 x_{k-1} + \phi_2 x_{k-2} + \ldots + \phi_p x_{k-p} + w_k - \theta_1 w_{k-1} - \ldots - \theta_q w_{k-q} \tag{7.12}$$

where the ϕ_i are autoregressive parameters which describe the effect of unit changes in x_{k-i} on x_k, and the θ_i are moving average parameters which describe the effect of past errors on x_k.

To be able to specify the signal process completely, using such a model, it is clearly necessary to be able to estimate the values ϕ_i and θ_i for as many terms as are required to yield the desired accuracy. Many files examined using preliminary analysis techniques, such as the $\alpha\beta$ tracker or running average, appeared to reveal signal processes which varied from stationary sequences (in the sense that little trend was evident) to those

which exhibited trend over periods both short and comparable to the duration of the measurement. The presence of significant temporal correlation, however, was always evident in the sense that none of the estimates appeared to be converging to the single valued estimate, characteristic of a stationary, white noise sequence, when filters were applied having characteristic time constants shorter than the duration of the measurement.

Much of the validity, in fact, of applying techniques such as the $\alpha\beta$ tracker and running average depend on selecting the responsivity such that, on the one hand, the responsivity is not too sluggish to reveal changes in absorption of a duration short compared with the measurement duration but, on the other hand, having a responsivity long compared to the temporal correlation expected of the measurement noise. Some guidance as the magnitude of the lower limit was available in the earlier work of Menyuk and Killinger [4]. For the running average, for example, 200 samples were used to form every estimate; 100 of these were samples prior to the estimate point and 100 after the estimate point. Since the samples were collected at approximately 10 Hz, 200 samples corresponds to a characteristic time constant of approximately 20 seconds. This is clearly short compared to the duration of the measurement but long compared to the atmospheric decorrelation time of 1 or 2 seconds, expected for speckle phenomena.

The remainder of this section is devoted to the application of autocorrelation techniques which are capable of providing some quantitative information on the composite signal, particularly with regard to its possible components. The "composite signal" is defined here to be that signal which results purely from variations in atmospheric absorption, combined with either multiplicative or additive measurement noise result-

ing from speckle. Additive speckle measurement noise is only available once the data has been log transformed. The additive noise case, however, is dealt with exclusively from now on due to its suitability for applying the methods of optimal estimation.

7.2.3.1 Log Power Data

The analysis used throughout Section 7.2.3 applies the "Univariate Stochastic Model Identification" (USID) technique outlined by Box and Jenkins in "Time Series Analysis", part V, algorithm 1 [7] and discussed in Section 6.3.2. A single Fortran program, PUSID.FOR, encoded and developed by Dr. B. J. Rye of Hull University, Department of Applied Physics, incorporated all the algorithms discussed here. Provision for the processing of seasonal data was not included in the program since periodic components were not expected to be a feature of the data. This data is assumed to exist as normalised powers in two sequential access files, PA.IN and PB.IN, corresponding to the two wavelength channels.

The program then provides the option of performing both an autocorrelation and partial autocorrelation on the data in this basic state or after several transformation options. These options are:

(1) PA.IN, PB.IN - no transformation

(2) Ln(PA.IN), Ln(PB.IN)

(3) Ln(PB.IN/PA.IN)

If the log transformations are selected, then a further option exists for including a "transformation shift" to move the data sequence away from zero or negative numbers.

Having selected the type of transformation, the degree of non-seasonal differencing may then be specified and finally, there is a further option of "pre-filtering" the data. If required, either "block averaging" may be used which is equivalent to applying the segmental average of Section 6.1.2, or else the data sequence may be sampled at some specified interval.

Figures 7.8(a) to 7.8(f) illustrate the results of applying PUSID to the two data files originally examined in Section 7.2.2. Each of the figures displays two curves; the solid line is the autocorrelation function (acf) of the log transformed normalised power data and the broken line is the corresponding partial autocorrelation function (pacf). Most figures are presented in pairs with channel A data uppermost and channel B data below (normalised powers are denoted PA and PB). Examining the data from file SIFT5.026 first, (Figures 7.8(a) and 7.8(b)), it is apparent that both the acf and pacf for both channels are identical, as would be expected for a measurement conducted using the same wavelength in each channel. Non-seasonal differencing has not been applied in either of these two examples and the data has not been subjected to any pre-filtering.

To facilitate model identification from plots such as these, a set of reference acf and pacf plots have been reproduced from Vandaele's book "Applied Time Series and Box Jenkins Models" [8], and are illustrated in Figure 7.9. The examples shown are based on the autoregressive model, AR(p), the moving average model, MA(q), and the synthesis of these two, the autoregressive moving average model, ARMA(p,q) for the cases p,q = 0,1,2 (see Section 6.3.2). Various values of the AR and MA parameters, ϕ and θ, are used and the examples culminate in a table (Figure 7.9(t)) summarising the salient features of these processes.

Comparing Figures 7.8(a) and 7.8(b) with this information it is evident that, of the examples shown, the closest approximations are represented by Figures 7.9(d) and 7.9(q), corresponding to the AR (p = 2) and the ARMA (p = 1, q = 1) with $\phi_1 = 0.6$ and $\phi_2 = 0.2$ in the former case, and $\phi_1 = 0.6$, $\theta_1 = 0.2$ in the latter. For the purposes of this comparison, the magnitude of these parameters is of less significance than their sign. Each of the three cases possess acfs which are of the same sign and exhibit relatively rapid decays over the first few lags. Clearly, extra information is necessary to resolve which of these candidate processes is closest to the real situation and it is here that the pacf becomes relevant.

Pacfs summarise all the information contained in the acf of an AR process using a small number of non-zero statistics such that, for an AR(p) process, only p such statistics are necessary [8]. The pacf, therefore, really shows the number and relative importance of the various AR terms incorporated into a model. Furthermore, any ARMA(p,q) process can always be expressed as pure autoregressive model [8]. Consequently it is useful to interpret any features characteristic of a potential ARMA process as equivalent to those of an AR process or higher order. Pacfs illustrated in Figures 7.8(a) and (b) indicate that an AR process with $p \simeq 5$ would form a suitable model for the observed data.

These two figures depart from Vandaele's examples, however, in one important respect and this is the significant acf values existing beyond approximately 20 lags after the initial rapid decay. It will be shown later that such features are characteristic of AR processes which approach the random walk limit ($\phi \rightarrow 1$) and become non-stationary. Figures 7.8(a) and (b) in fact represent an extreme example of this significant finite acf at high lags, but as will also be seen later, the same feature tends

to be present in all the other data files examined, although often to a lesser degree. As an example, Figure 7.8(c) and (d) illustrate the acfs and pacfs for data file SIFT5.045 (a water vapour measurement). Again, the characteristic rapid decay is followed by a small but finite acf which persists out to 40 or 50 lags for this particular file. The pacf plot, however, indicates an AR model with less than ten significant terms, as was the case for the previous data file, SIFT5.026.

Menyuk and Killinger obtained similar plots to these for the data files they examined, although they did not apply the logarithm transformation to any of their data [4]. At an early stage in the analysis of Hull lidar data, the program PUSID was applied to single channel normalised power data, ratioed data and the logarithm of both of these but very little difference was observed between the transformed data and non-transformed data, despite the non-linearity of the log transformation operation.

Menyuk and Killinger attempted to fit analytical expressions to the autocorrelation function they obtained and essentially partitioned these functions into two regions. The first, characterised by the rapid decay in the acf, they found to be adequately described by a function of the form,

$$\rho_j = \exp(-j\alpha) \tag{7.13}$$

where ρ is the correlation coefficient, j denotes lag, and α is a constant characteristic of the decay rate. The second region characterised by a finite, almost constant acf out to greater lags was found to obey a logarithmically decreasing expression of the form,

$$\rho_j = A - B \ln(j) \tag{7.14}$$

and it is to this component that they attribute the failure of the variance to achieve an $N^{-\frac{1}{2}}$ dependence.

Autocorrelation and partial autocorrelation functions of the form discussed above can only be obtained for composite signal data (i.e. normalised power signal in the absence of speckle, combined with the speckle process) since, clearly, there is no way of separating the measurement noise from an underlying process in order to observe the auto-correlations of each. In forming a signal model for optimal estimation purposes certain assumptions, therefore, had to be made regarding which features of the acfs were plausibly attributable to the signal component and which to the speckle dominated measurement noise.

Menyuk and Killinger point out that "previous studies of the temporal correlation coefficients for backscattered CO_2 laser radiation through the atmosphere from a hard target have indicated that the atmosphere is effectively frozen for the order of 1 - 5 msec, with a significant drop in correlation by 50 - 100 msec." Lidar systems operating at 10 Hz, therefore, (as was the case with both the Hull University lidar system and the system used by Menyuk and Killinger) would exhibit acfs characteristic only of the "tail end" of the full autocorrelation function, the remainder of which is assumed to exist for interpulse delays in the 1 - 100 msec range. Short term decorrelation ($\simeq$ 1 second), therefore, is attributable to changes in the state of the atmospheric channel between laser shots and, for diffraction limited systems, is due primarily to "turbulence caused by thermal fluctuations in the atmosphere" [4].

As pointed out earlier, speckle induced variance arises due to changes, between shots, in the speckle pattern imaged at the detector. The speckle pattern formed at any instant at the receiver is dependent both on random phase changes induced by the atmosphere itself and also by shifts in the

position of the laser beam across the surface of a non co-operative target. Both can, however, originate with changes in the refractive state of the atmospheric channel. Therefore, since speckle induced variance is far greater than that expected for any signal process, it does not seem unreasonable to attribute the initial rapid decay in the acf to speckle measurement noise decorrelation. Beyond the 10 or 20 lags (≡ 1 or 2 seconds) characteristic of this "atmospheric decorrelation time" finite values of the autocorrelation function may then be assumed to be attributable to changes in the absorptive state of the atmosphere. If, for example, this absorptive state were to remain stationary throughout a measurement, the acf would be expected to decay quickly to zero after 10 or 20 lags. The $N^{-\frac{1}{2}}$ dependency for the variance of an average based on N values, but separated by at least the speckle decorrelation interval, would again be expected to hold.

Returning to Figure 7.8, (e) and (f) present the acf and pacf of data for which first order differencing has been applied after the data has been log transformed. Implicit in the application of any differencing is the suspicion that the data may in fact be more complex than an ARMA process and may require modelling using the more general form of an ARIMA (p,d,q) or "Autoregressive Integrated Moving Average" process, where d denotes the degree of differencing necessary to reduce an ARIMA process to an ARMA process. A comparison of Figure 7.8(e) and (f) with the examples of Figure 7.9, reveals that the first order moving average (7.9(j)) processes a very similar acf and pacf to those two figures. A question therefore arises as to which interpretation of the data is more relevant; the acf and pacf for d = 0 (Figures 7.8(a) to 7.8(d)), or those for d = 1.

Box and Jenkins advocate the use of differencing to render a time series stationary, at least with respect to its first and second statistical moments (mean and variance). First order differencing may be applied, for example, where it is assumed that the time series possess a linear trend. The advisability, however, of applying a particular degree of differencing depends not only on such factors as the polynomial order of any trend believed to be present, but also on the signal to measurement noise ratio. Although something of analytical significance may be gained by imposing stationarity on the signal component via differencing, this same differencing operation is also unavoidably applied to the measurement noise. It may therefore result in further degradation of the information content of the composite signal as the spectral energy density function of the measurement noise is shifted to higher frequencies and higher powers.

Without exception, however, all of the lidar data analysed by applying first order differencing, produced almost identical acf and pacf characteristics to those illustrated in Figures 7.8(e) and 7.8(f), irrespective of whether the data was in normalised power form, power ratio or log transformations of these. Simulations will be used later (Section 7.2.3.3) to demonstrate that this characteristic MA (1) process is easily generated using very simple composite signal models, emplying both white and coloured speckle sequences.

The remainder of the figures identified as 7.8(g) to 7.8(r) present the acf and pacf results for log power data recorded in both channels for each of the data files considered here (listed in Table 7.1). None of these have been subjected to the differencing operation. Varying rates of decay are evident in the initial $\sim$ 10 lags and this may be attributable to varying turbulence conditions during the measurement; the greater the

strength of the turbulence, the more the decorrelation expected in the atmospheric channel between laser shots. Beyond this region of relatively rapid decay, the autocorrelation function behaviour exhibits either very slow decay or appears to maintain some constant non-zero value over the remainder of the lag range indicated. Also, the pacf tends to indicate that approximately the same number of autoregressive terms are involved in each case, but with some differences in the relative strengths between these contributing terms.

7.2.3.2 Log Ratio Data

Having examined, separately, the autocorrelation characteristics of the log power measured in each wavelength channel, Figure 7.10(a) to 7.10(i) now present the acf and the pacf characteristics of the same files but this time using log ratio data. Each of the figures in 7.10 may be compared on a file by file basis with their equivalents, for log power data, in Figure 7.8.

An indication of the significance of correlated speckle is provided by a comparison between the first example, Figure 7.10(a), and the earlier Figures 7.8(a) and (b). Both of the latter acf plots exhibit high autocorrelation values above and below that transition region identified earlier where the rate of decay changes (most often near lag = 10 but sometimes out to lag = 20, as with SIFT5.026 for example). Their log ratio, however, presents a considerably reduced acf decay over this same initial lag region and virtually no correlation beyond 20 lags.

For this file, therefore, samples based on the log ratio (or ratio) are nearly independent and are consequently of greater statistical significance for variance reduction purposes in any estimation algorithm.

Menyuk and Killinger's data demonstrates identical behaviour for the ratio of normalised powers [4].

A valuable characteristic of autocorrelation functions, therefore, is the indications they provide of sample independence. Unless an estimation algorithm is specifically designed to cope with temporally correlated data, it may be advisable to sample the data, or take measurements, at intervals beyond which the data is known to be significantly correlated. For the lidar data presented here, and particularly log-power data, a sample interval of approximately one in ten is often appropriate if successive samples must be independent. Although the system was operating at 10 Hz when these files were collected, a requirement for complete sample independence would limit the sample rate to a maximum useful value of 1 Hz if similar measurements were to be conducted again. Fortunately, however, techniques exist in optimal estimation (discussed in Section 6.3) for modelling correlated measurement noise, which makes it possible to collect useful samples at much faster rates. These techniques are applied, and the results analysed, in Section 7.3.

The next figure, 7.10(b), presents the acf and pacf for a data file using water vapour measurement lines R20 and R18 (SIFT5.045). Since the data sequences are much less correlated for this file (correlation coefficient $\simeq$ 0.1 as opposed to 0.9 for the previous file, SIFT5.026 - see Table 5.1), its log ratio acf values, in the lag region less than or equal to 10, are only slightly less than those illustrated in Figures 7.8(c) and 7.8(d) for log power data from the same file. Beyond lag = 10, the acf is similar in magnitude and appears to decay very slowly. With first order differencing applied, the acf and pacf for file SIFT5.045 are as illustrated in Figure 7.10(c). As can be seen, these display characteristics almost identical to those illustrated in Figures 7.8(e) and 7.8(f) for

log power data. Further comments on this type of autocorrelation result will be reserved until Section 7.2.3.3.

Remaining figures, 7.10(d) to 7.10(i), present the log ratio acf and pacf plots for the same collection of files as examined in previous sections, all of which used dissimilar wavelengths in the two lidar channels (wavelengths used are identified in Table 7.1), and for which no differencing has been applied. The use of dissimilar wavelengths results in correlation coefficients between channels not exceeding a maximum of 0.22 (for SIFT5.046 - see Table 5.1). Consequently, most of the log ratio correlation functions appear very similar to those illustrated in Figure 7.8, for the log powers from the same data file, as observed in either channel A, channel B,or both.

7.2.3.3 Simulated Data

The series of autocorrelation and partial autocorrelation plots discussed below are all based on the simulated data files presented earlier. In the first example, Figure 7.11(a), the autocorrelation characteristics are presented for the simulated data file which additively combines a constant signal value of -1.5 with a white (uncorrelated) Chi-square sequence (M = 3, variance = 0.25) representing speckle measurement noise. Both the acf and pacf are coincident at a constant value of zero for all lags (allowing for noise in the autocorrelation function estimate), as expected for a sequence which is, essentially, just white noise. The result is, however, confirmed in Figure 7.9(a) and provides a check on the implementation of the algorithms used to process the data.

The following figure, 7.11(b), is for the simulated data file which, again, used a constant signal (= -1.5) but is combined additively with

coloured (correlated) speckle measurement noise. As indicated earlier, the decorrelation parameter was specified to reduce the acf to 0.1 after 5 lags (or every five "shots") and this figure demonstrates that this is the case.

The pacf reveals the presence of a single autoregressive term, as expected, since a single decorrelation parameter was used, and both plots are in agreement with the examples of Figure 7.9(b). Although larger in magnitude, the acf decay over the first ten lags is of the same general form as that exhibited by those real data files considered earlier, in which temporal correlation was evident in this same lag region. It therefore will serve as a worst case example for suspected speckle correlation.

Figure 7.11(c) illustrates the acf and pacf for the first order difference of this simulated data file. Similarities between this simulated example and the equivalent plots for real data become apparent when this figure is compared with the previous Figures 7.8(e) and 7.8(f) for log power data and 7.10(c) for log ratio data. If the simulated data file for the correlated speckle case is now considered, in which speckle was again superimposed on a constant signal of -1.5, the autocorrelation characteristics are as shown in Figure 7.11(d).

Although a similar dependence on lag is revealed, the magnitude of the autocorrelation terms at lower lag values (< 10) are significantly less than those shown by the real data, and by the simulated example for white speckle. Furthermore the acf for the simulated data exhibits a gradual decay, characteristic of an ARMA process of order 1,1 or higher, in which each of the coefficients is of positive sign (cf with Vandaele's example, 7.9(p)). By contrast, for all of the real data files examined using first order differencing, the presence of only a single significant term in the acf was revealed in each and every case, and this occurs at

lag = 1. Such a result, however, is not inconsistent with the properties of time series and their autocorrelation characteristics. As will be explained below, the observed results can be highly dependent on parameters such as the autoregressive coefficient, ϕ, even when relatively simple models such as the AR(1) are considered.

Modelling correlated speckle as a first order exponential decay is equivalent to stating that the process is an autoregressive process of order 1. This equivalence was indicated in Section 6.3 and is reproduced here for convenience. A first order exponentially correlated sequence takes the discrete form

$$S_k = e^{-\beta\Delta t} S_{k-1} + a_{k-1} \tag{7.15}$$

where β is the reciprocal of some time period characteristic of the decay rate, Δt represents the temporal lag interval, S denotes the speckle sequence, and a is zero mean gaussian white noise sequence.

An AR(1) series takes the form,

$$S_k = \phi_1 S_{k-1} + a_{k-1} \tag{7.16}$$

which, for fixed values of β and Δt, is clearly identical to (a) and has an autocorrelation function defined by (see Section 6.3)

$$\rho_k = \phi_1^k \tag{7.17}$$

For the simplified case in which a speckle sequence is combined additively with a constant signal (= -1.5 for the simulated data file considered here), the application of differencing will only affect the speckle sequence since the first difference of a constant signal will be zero. Therefore, assuming the speckle sequence is of the generalised AR(1) form given above, performing a first order difference yields,

$$s_k - s_{k-1} = \phi_1(s_{k-1} - s_{k-2}) + a_{k-1} - a_{k-2} \tag{7.18}$$

or $$\omega_k = \phi_1 \omega_{k-1} + a_{k-1} - a_{k-2}$$

where ω represents the differenced terms.

This can be cast into the general form of an ARMA(1,1) model;

$$\omega_k = \phi_1 \omega_{k-1} + a_{k-1} - \theta_1 a_{k-2} \tag{7.19}$$

where $\theta_1 = 1.0$. For a simulated white speckle sequence, $\phi_1 = 0.0$, since successive values must be statistically independent, and the model reduces to

$$\omega_k = a_{k-1} - a_{k-2} \tag{7.20}$$

which is the MA(1) process observed in the autocorrelation plots of Figure 7.11(c), and is also characteristic of real data.

In Section 6.3.2 the autocorrelations of an ARMA(1,1) model were shown to be given by

$$\rho_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 + \theta_1^2 - 2\theta_1\phi_1} \tag{7.21}$$

$$\rho_k = \phi_1 \rho_{k-1}, \quad k \geq 2$$

but since $\theta_1 = 1.0$, this reduces to

$$\rho_1 = \frac{(\phi_1 - 1)}{2} \tag{7.22}$$

$$\rho_k = \phi_1 \rho_{k-1}, \quad k \geq 2$$

An MA(1) process ($\phi_1 = 0$, $\theta_1 = 1$) therefore will have a single term $\rho_1 = -0.5$ at lag $k = 1$, but all other autocorrelations for $k \geq 2$ will

be zero. This is clearly indicated in Figure 7.11(c).

If the simulated data file, which additively combines the constant signal with correlated speckle is now considered, the required value of ϕ_1 can be derived from (7.17) since this sequence was specified to decay to a correlation coefficient of 0.1 after 5 lags. In other words, since

$$\rho_k = \rho_5 = 0.1 = \phi_1^5$$

ϕ_1 is then found to have a value of 0.63. After first order differencing has been applied, the sequence is described by the model (7.18) for which $\theta_1 = 1.0$, as before. The correlation coefficient at the first lag, found using (7.22) above, is

$$\rho_1 = (0.63 - 1.0)/2 = -0.185$$

with the correlation coefficient at subsequent lags (≥ 2) decaying as also shown by (7.22). This corresponds to the result in Figure 7.11(d).

A value of $\phi_1 = 0.63$, as used in this simulated data file, yields an exaggerated acf decay when compared with most of the equivalent plots for real data (compare Figure 7.11(b) with those of Figures 7.8 and 7.10). In general, as ϕ_1 decreases towards zero, (7.22) above indicates that ρ_1, will approach -0.5 for the first order differences of data modelled on an AR(1) process, but that ρ_2 and successive correlations will rapidly decrease to zero. Therefore, simulated data exhibiting lower acf terms over the first few lags can be expected to produce autocorrelation characteristics similar to those observed for real data, both in the absence of differencing and with first order differences applied.

If real data is now considered then, taking as an example the log-powers of channel A data from file SIFT5.045, Figure 7.8(c) gives $\rho_1 = \phi_1 \simeq 0.28$. Expression (7.22) yields,

for the differenced data, $\rho_1 = (0.28 - 1)/2 = -0.36$ and $\rho_2 = -0.36 \times 0.28 = 0.1$ (autocorrelation terms at higher lags will be negligible). If Figure 7.8(f) is now examined, the actual value obtained for ρ_1 is nearer to 0.43 and ρ_2 is indistinguishable from the noise existing in these acf estimates. This noise, however, is equivalent to an error of approximately ±0.05 in each correlation term (as estimated from the figures).

Consequently, even though some discrepancy exists between the values predicted assuming an AR(1) process, and those derived from real data, most of the discrepancy may in fact be due to errors in the acf estimate (and attributable to the finite extent of the data sequence). Furthermore, some component of the autocorrelation function over the first four lags is likely to be due to correlation in those absorptive processes, characteristic of the atmospheric state, which underlie speckle measurement noise. The first order autoregressive process of equation (7.16) is therefore considered a reasonable model for approximating the various rates of decay observed in the autocorrelation functions of all real data files over, approximately, the first ten lags.

So far, only two of the simulated data files, introduced in earlier sections of this chapter, have been considered here; both employed the artificially simple assumption that the absorptive state of the atmosphere remained constant throughout an equivalent measurement sequence of 10,000 "shots". One version, however, was additively combined with uncorrelated (white) speckle and the other with correlated (coloured) speckle. The other two simulated data files, also introduced earlier, replaced the constant signal with a single random walk process but additively combined this with the same two versions of the speckle sequence as used in the constant signal simulations.

A random walk is a special case of the AR(1) process for which $\phi_1 = 1$. It is a non-stationary time series which, for finite realisations, will possess statistical moments dependent on the length of the series, but in the limit in which the series becomes infinite these moments will, of course, also become infinite. For all values of $|\phi_1|$ less than unity, an AR(1) series is stationary in the sense that at least the first two statistical moments, (mean and variance) have constant finite values for all realisations of the series. As $|\phi_1|$ approaches unity, however, an AR(1) sequence becomes increasingly like a random walk and therefore develops the useful potential for modelling trend.

For values of $|\phi_1|$ very close to unity (for example those existing in the interval $\phi_1 = \pm 1.0 \pm 0.001$) therefore, an AR(1) series is capable of modelling trend and if $|\phi_1| < 1$ it will also possess constant, finite values of mean and variance. If the signal process is denoted by x then by representing the AR(1) series in form

$$x_k = \phi_1 x_{k-1} + w_{k-1} \tag{7.23}$$

where w is a zero mean gaussian noise sequence, the variance of such a series can be presented as (see section 6.3.2)

$$\sigma_x^2 = \frac{\sigma_w^2}{(1 - \phi_1^2)} \tag{7.24}$$

Figure 7.7(a) illustrated the random walk sequence used as the signal in the above mentioned simulated data files. For this simulation the value of ϕ_1 used was, in fact, unity and the variance, $\sigma_w^2 = 10^{-5}$. Selecting values of ϕ_1 between 0.999 and 1.0 and using the same initial random number seeds will generate a family of plots, all with the same "shape" but compressed in the ordinate axis as ϕ_1 decreases. Different random number seeds produce signals of differing "shapes", each manifest-

ing different extremes of trend. A value of ϕ_1 equal to unity was adopted in the simulated data files because the same value is assumed for the AR(1) signal model, later to be used when applying optimal estimation techniques.

The two Figures, 7.11(e) and 7.11(f), present the acf and pacf results for a pure random walk sequence in the absence of speckle (no differencing applied). A value of $\sigma_w^2 = 5 \times 10^{-5}$ was used to generate this particular sequence. As can be seen, the autocorrelation function displays a gradual decay from unity characteristic of a finite random walk. For an infinite series, the acf would presumably reveal a constant value of unity for all lags at all finite, non-zero values of the driving noise variance σ_w^2. The pacf plot illustrates the presence of a single autoregressive term at lag = 1, as expected (the nearest example to this amongst Vandaele's figures is illustrated in Figure 7.9(b)). To complete the autocorrelation characteristics for the random walk sequence, plots for the first order differences have been included (Figures 7.11(g) and 7.11(h)) which confirm the somewhat trivial result that the acf and pacf of the first order difference of a random walk are zero for all lags.

An additive combination of the random walk and the white speckle noise sequence (previously illustrated in Figure 7.7(b)) has an acf and pacf as presented in Figure 7.11(i). Of immediate significance here, is the constant non-zero value of the autocorrelation function persisting over the entire lag range included in the figure.

Since, for many of the real data files examined using PUSID, persistent non-zero, acf values of similar magnitude have been observed in the lag region beyond the initial rapid acf decay, the composite simulation employing the random walk and white speckle appears to be manifesting at least one of the required characteristics of real data. The pacf

on the other hand would seem to indicate the presence of a series of autoregressive terms which decay in significance as the order of each term increases.

If the same random walk is now additively combined with correlated speckle noise (as illustrated previously in Figure 7.7(c)) then the resulting autocorrelation characteristics are as presented in Figure 7.11(j). Here, the decaying acf in the initial lag region, due to the presence of correlated speckle, merges with the acf component beyond lag $\simeq$ 10 due to the random walk signal model.

The effect of the random walk contribution is further emphasised by comparing Figure 7.11(j) with 7.11(b) which illustrated the results for correlated speckle with a constant signal. Both plots appear identical apart from the shift in the acf occurring at all lags due to the presence of the random walk component. Some extra noise, however, is also evident in each of these acf estimates due, presumably, to certain features common to both data files; the finite extent of the data set combined with significant correlation between samples is a possible explanation. Both examples also possess almost identical partial autocorrelation functions, and in general are consistent with an AR(1) process dominated in different lag regions by either correlated speckle noise or the random walk signal model.

Finally, by applying first order differencing to the simulated data which combines the random walk with correlated speckle, the autocorrelation plots of Figures 7.11(k) and 7.11(l) are obtained. These are almost identical to Figures 7.11(c) and 7.11(d), which correspond to the constant signal case with additive white speckle noise and coloured speckle noise respectively. This is expected since the first order differences of a random walk process will produce the zero mean gaussian noise sequence,w,

used to drive the random walk (see expression (7.23)) and such a sequence contributes nothing to the autocorrelation or partial autocorrelation functions. Both are therefore entirely due to the simulated speckle sequences.

In summary, therefore, the first order autoregressive series appears to be an adequate approximation for modelling those processes occurring in the real data, at least as far as their autocorrelation characteristics are concerned. Although more accurate models could probably be derived by pursuing, further, the Box-Jenkins approach to time series analysis or, alternatively, the methods advocated by Peter Young [9], these models will almost certainly be more complex. For the purposes of applying the methods of optimal estimation, relatively simple models were sought which would facilitate an initial implementation of those techniques of particular relevance to the processing of lidar data. These techniques and the success with which they were applied to AR(1) models, using the real and simulated data files analysed above, will form the subject of discussion in the following section.

7.3 The Application of Optimal Estimation Methods

Section 6.3 presented, in general terms, a selection of techniques collectively entitled "Optimal Estimation" of which Kalman filtering theory forms the topic of central importance for the analysis conducted here. However, because this subject area is so wide ranging in existing and potential applications, it is necessary to point out that the other topics included there form a suite of techniques selected in an attempt to "customise" the Kalman filter for the estimation of lidar data. This section will consider specific implementations of these methods and their application to lidar data. Model identification has already been dealt

with in some detail above but, in practice, is generally recognised that the process of identification also includes repeated trial applications of candidate models in an iterative approach which seeks to optimise the models, and hence the filter. In this sense, therefore, the following analysis represents a first iteration.

7.3.1 Implementation of the Estimation Algorithms

A summary of the discrete Kalman filter equations has already been presented in vector/matrix notation in Chapter 6, but they are reproduced here in scalar form to facilitate further consideration:

System Model: $$x_k = \phi_k x_{k-1} + w_{k-1}\ ;\ w_k = N(0,q_k) \tag{7.25}$$

Measurement Model: $$z_k = h_k x_k + v_k\ ;\ v_k = N(0,r_k) \tag{7.26}$$

$$E[w_k v_j] = 0 \text{ for all } j,k \tag{7.27}$$

State Estimate Extrapolation: $$\hat{x}_k(-) = \phi_{k-1} x_{k-1}(+) \tag{7.28}$$

Error Covariance Extrapolation: $$p_k(-) = \phi_{k-1}^2\ p_{k-1}(+) + q_{k-1} \tag{7.29}$$

State Estimate Update: $$\hat{x}_k(+) = \hat{x}_k(-) + K_k[z_k - h_k \hat{x}_k(-)] \tag{7.30}$$

Error Covariance Update: $$p_k(+) = [1 - K_k h_k]\ p_k(-) \tag{7.31}$$

Kalman Gain: $$K_k = p_k(-)\ h_k[h_k^2\ p_k(-) + r_k]^{-1} \tag{7.32}$$

Autocorrelation techniques applied in Section 7.2.3 revealed that the first order autoregressive model AR(1), of equation (7.13) was a reasonable approximation to the signal process for values of ϕ close to unity. Reducing the discrete Kalman filter equations to the scalar form above

is, in fact, tending towards the assumption that an AR(1) process is an adequate signal model. The first order autoregressive process also appears to be a reasonable approximation for describing correlated speckle sequences.

As a further note, therefore, on the significance of autocorrelation functions, relevant to both signal and measurement models, Gelb points out that "as a practical fact, most often all we know about the characterisation of a given random process is its autocorrelation function. But there always exists a gaussian random process possessing the same autocorrelation function; we therefore might as well assume that the given random process is itself gaussian. That is, the two processes are indistinguishable from the standpoint of the amount of knowledge postulated" [ref. 10, p. 105].

The simulated data file examined in Section 7.2.3 used $\phi = 1.0$ but, as was pointed out, values of ϕ confined to $\pm 1.0 \pm 0.001$ will generate a family of time series, all exhibiting similar random trend features if the same random number seeds are used. For values of $|\phi| < 1$, the series (7.23) will also possess a mean of zero and a variance defined by (7.24). The system model of expression (7.25) denotes the autoregressive coefficient as a variable ϕ_k, dependent on the recursive index, k. If ϕ_k is to be regarded as a parameter to be estimated then at least two possible methods of deriving it seem to be available:

(1) Individual autocorrelation terms in the acf and pacf plots discussed earlier can be examined, as was attempted in Section 7.2.3.3. Estimates based on these will be subject to error in the acf/pacf estimates and will yield either a single value, $\phi_k = \phi$, based on the entire measurement sequence, or multiple values based on subsets of it.

(2) The state variable, x, can be augmented to a vector which also contains ϕ or ϕ_k as an additional term to be estimated. This will also be subject to some estimation error.

However, any AR(1) system model employed within a Kalman filter algorithm to estimate data exhibiting trend, must have a value of ϕ so close to unity that an attempt to estimate ϕ will be in error by at least $1 - |\phi|$. Without an in-depth analysis to identify methods for confining errors in the estimate of ϕ to within this limit, therefore, it would seem a reasonable approximation to set $\phi_k = \phi = 1$.

The system model adopted is therefore the random walk. To facilitate the implementation of techniques for deriving system and measurement noise variances, both q and the measurement noise variance, r, are assumed to remain constant throughout the measurement, so that $q_k = q$ and $r_k = r$. In some future development of these algorithms, which is beyond the scope of the present work, adaptive filtering techniques could be investigated which would permit both q and r to be recursively estimated during the measurement. As indicated by the compact notation N(0,q) and N(0,r), used in expression (7.25) and (7.26), the two noise sequences, w and v, are also assumed to have normal (gaussian) distributions with zero means.

As a further simplification, h_k in the measurement model (7.26), can be set to a constant value of unity ($h_k = h = 1$) so that the measurement is modelled as a simple summation of the log transformed signal (denoted x, and which may apply to either log power or log ratio data) and the measurement noise, v, which is assumed to be dominated by speckle. Expressions (7.25) through to (7.32) therefore become

System Model: $$x_k = x_{k-1} + w_{k-1} \tag{7.33}$$

Measurement Model: $$z_k = x_k + v_k \tag{7.34}$$

State Estimate Extrapolation: $$\hat{x}_k(-) = \hat{x}_{k-1}(+) \tag{7.35}$$

Error Covariance Extrapolation: $$p_k(-) = p_k(+) + q \tag{7.36}$$

State Estimate Update: $$\hat{x}_k(+) = \hat{x}_k(-) + K[z_k - \hat{x}_k(-)] \tag{7.37}$$

Error Covariance Update: $$p_k(+) = [1 - K_k]\, p_k(-) \tag{7.38}$$

Kalman Gain: $$K_k = p_k(-)/(p_k(-) + r) \tag{7.39}$$

If (7.37) is compared with equation (6.55) for the first order recursive filter, it is apparent that the two estimators are identical in form. However, the version derived using Kalman filter theory has a variable gain, K, which is adjusted as the filter index, k, increases according to expression (r), until the steady condition is reached in which p and K do not change with further increases in k. Steady state values of p and K can be obtained from the "matrix Riccati equation" which, in scalar form, becomes (see Section 6.3.1)

$$\dot{p} = 2fp + g^2q - p^2/r \tag{7.40}$$

and also from the continuous system model equation which, in a generalised scalar form is (see Section 6.3.1)

$$\dot{x} = fx(t) + gw(t) \tag{7.41}$$

where $\dot{p}$ and $\dot{x}$ denote the derivatives of p and x with respect to time.

For the random walk, f = 0 and g = 1, so that (7.40) becomes

$$\dot{p} = q - p^2/r \tag{7.42}$$

but in the steady state condition, $\dot{p} = 0$, so that

$$p_{\infty} = \sqrt{qr} \tag{7.43}$$

where p_{∞} denotes the steady state error covariance. Also, since the gain matrix, K(t), for the continuous Kalman filter equations is, in general terms, given by (see Section 6.3.1),

$$K(t) = P(t)\ H^{T}\ R^{-1}(t) \tag{7.44}$$

then reducing to the scalar form and replacing P(t) with the steady state error covariance, p_{∞}, yields the steady state gain,

$$K_{\infty} = p_{\infty}/r = \sqrt{(q/r)} \tag{7.45}$$

Once q and r are determined, therefore, an estimate of the steady state error covariance, p_{∞},becomes available which may then be used to quantify the error in the estimate of the state variable, x.

The algorithm specified by the set of expressions (7.33) to (7.39) forms a first stage implementation of the Kalman filter. It is unable to cope, however, with correlated measurement noise since this is not built into the measurement model, (7.34). Data used with this algorithm must therefore be sampled from the original data set at a lag interval sufficient to guarantee that speckle correlation is negligible. For most of the data files discussed here, a sampling interval of 1 in 10 is sufficient but there are at least two files for which sampling of 1 in 20 must be considered. Guidance as to the correct sampling interval is, of course, available from the autocorrelation functions previously estimated for each file.

Built into this first stage Kalman filter is a version of Mehra's algorithm for estimating q and r [11]. The essential elements of this algorithm were discussed in Section 6.3 and its application requires an iterative procedure which may be summarised as follows:

(1) Sample the data set at an interval greater than or equal to the speckle decorrelation interval. This subset of the original data file forms the input to the first stage Kalman filter.

(2) Form rough estimates of q and r using the sampled data.

(3) Run the first stage filter on the sampled data and form the autocorrelation function (ρ_ℓ) of the innovation sequence, $\nu_k = z_k - x_{k-1}$.

(4) Estimate new values of q and r.

(5) Perform optimisation (whiteness) test on the innovation sequence.

(6) If the innovation sequence is coloured, use the new values of q and r, and re-filter the data. If the innovation sequence is white, proceed to the next stage.

The autocorrelation function, $\hat{\rho}_\ell$, of the innovation sequence is estimated for 20 lags ($\ell = 1,2,..20$) during each run of the Kalman filter. An optimisation test is then conducted by:

(a) Counting the number of autocorrelation terms, Iacf, which are outside the 95% confidence limits, $\pm(1.96/N^{-\frac{1}{2}})$, defined by Mehra (see Section 6.3.3). Ideally, this number should be ≤ 1 for the sequence to be white.

(b) Counting the number of autocorrelation terms, Ineg, which are less than zero - this will reflect the amount of bias present in autocorrelation function and for optimum results

will be near 10.

N is the number of innovation terms, ν_i, over which the autocorrelation is estimated and is equivalent to the number of values in the sampled data set, less the first 10% during which the filter is allowed to reach the steady state.

To estimate new values of q and r when the whiteness test fails, use is made of Mehra's equation (see Section 6.3.3),

$$\hat{M}\hat{H}^T = K\,\hat{C}_o + A^* \begin{bmatrix} \hat{C}_1 \\ \hat{C}_2 \\ \vdots \\ \hat{C}_n \end{bmatrix} \tag{7.46}$$

where

$$A = \begin{bmatrix} H\Phi \\ H\Phi(I - KH)\Phi \\ \vdots \\ H[\Phi(I - KH)]^{n-1}\Phi \end{bmatrix} \tag{7.47}$$

$\hat{M}$ is an estimate of the steady state error covariance matrix, and K is the sub-optimal gain matrix. Since n is the state vector dimension and is equal to unity for the scalar version of the filter, both the matrix A and its pseudo inverse A* also become unity because,

$$A = H\Phi = h = 1$$

This simplification reduces expression (7.41) to

$$\hat{m} = K\,\hat{C}_o + \hat{C}_1 \tag{7.48}$$

An estimate of r is found using another of Mehra's equations

$$\hat{R} = \hat{C}_o - H(\hat{M}\hat{H}^T) \tag{7.49}$$

which, for the scalar case, reduces to

$$\hat{r} = \hat{C}_o - \hat{m}$$

$$= \hat{C}_o (1 - K) - \hat{C}_1 \tag{7.50}$$

The system noise variance, $\hat{q}$, can now be obtained by solving (6.157) and (6.158). In scalar form this expression reduces to the solution,

$$\hat{q} = -\Omega = 2K\hat{m} - K^2 \hat{C}_o$$

$$= 2K(K \hat{C}_o + \hat{C}_1) - K^2 \hat{C}_o$$

$$= K^2 \hat{C}_o + 2K C_1 \tag{7.51}$$

New estimates of q and r are repeatedly obtained by re-running the first stage Kalman filter with the previous values of q and r until the optimisation test indicates that the innovation sequence is white.

The next stage in filtering employs a version of the Kalman filter algorithm developed by Bryson and Hendrikson (see Section 6.3), for estimating a signal in the presence of sequentially correlated noise [12]. This should permit all values in the original data set to be used in forming an estimate. The measurement noise variance, r, will remain the same regardless of any sampling used but the system noise, $\hat{q}$, must be scaled down since the estimate derived above is based on a random walk model which is sampled at some multiple of the original sampling period used at the time of the measurement. Since the random walk possesses a variance which increases linearly with time, an appropriate scaling relationship is

$$q_1 = \frac{q_n}{n} \tag{7.52}$$

where q_n is the estimate, $\hat{q}$, formed by sampling every nth value from the original data set.

Bryson and Hendrikson's algorithm was presented in Section 6.3 and is reproduced here in its scalar form for the random walk signal model ($\Phi = \phi = 1$, $H = h = 1$). Using their notation, the system and measurement models become:

$$x_{k+1} = x_k + w_k , \qquad w = (0, q_1) \tag{7.53}$$

$$z_k = x_k + \varepsilon_k \tag{7.54}$$

where $$\varepsilon_{k+1} = \Psi\varepsilon_k + u_k , \qquad u = (0, \bar{q}_1) \tag{7.55}$$

and $$\zeta_k = z_{k+1} - \Psi z_k \tag{7.56}$$

is the transformed measurement. The filtering algorithm is then,

$$S = QH^T = q_1 h = q_1 \tag{7.57}$$

$$R = \bar{Q} + HQH^T = \bar{q}_1 + h^2 q_1 = \bar{q}_1 + q_1 \tag{7.58}$$

$$D = SR^{-1} = q_1/(\bar{q}_1 + q_1) = d \tag{7.59}$$

$$H^r = H\Phi - \Psi H = h\phi - \Psi h = (1 - \Psi) = h^r \tag{7.60}$$

$$K_{k-1} = m_{k-1}\, h^r/[(h^r)^2\, m_{k-1} + r] \tag{7.61}$$

$$P_{k-1} = (1 - K_{k-1}\, h^r)^2\, m_{k-1} + K_{k-1}^2\, r \tag{7.62}$$

$$m_k = (1 - dh^r)^2\, p_{k-1} + q_1 - d^2 r \tag{7.63}$$

where $p_{k/k} = m_k$ and $p_{k-1/k} = p_{k-1}$.

Use of subscripting of the form $p_{k-1/k}$ implies that the (k - 1)th estimate of p is based on all samples up to, and including, the kth sample. The filter then becomes

$$\hat{x}_{k/k} = x_{k-1/k-1} + [d + (1 - dh^T) K_{k-1}](\zeta_{k-1} - h^T \hat{x}_{k-1/k-1}) \tag{7.64}$$

The correlated measurement noise, ε, is clearly of the AR(1) form discussed earlier. However, both the autoregressive coefficient, ψ, and the variance, $\bar{q}_1$, of the white noise sequence, u, must be determined before this algorithm can be used. Providing a method for evaluating ψ can be found, $\bar{q}_1$ may be obtained by comparing (7.55) with (7.23) and using expression (7.24) to give,

$$\bar{q}_1 = \sigma_u^2 = (1 - \psi^2) \sigma_\varepsilon^2 \tag{7.65}$$

Usually, one of the first techniques to be considered when an unknown such as ψ must be estimated, is the augmentation of the state vector. Gelb, however, demonstrates a fundamental problem with this approach which results in the measurement noise matrix, R, becoming singular for the continuous Kalman filter, or the covariance update possibly becoming "ill conditioned" ($P_{k+1}(-) \simeq P_k(+)$ for small Q and $\Phi \simeq I$) in the discrete version of the filter [ref. 10, pages 133-136].

The method employed here, therefore, makes use of the knowledge that expression (7.55) is a discrete representation of an exponentially correlated process having the alternative form,

$$\varepsilon_{k+1} = e^{-\beta|\tau|} \varepsilon_k + u_k \tag{7.66}$$

and possessing an autocovariance function described by

$$\lambda_\varepsilon(\tau) = \sigma_\varepsilon^2 \, e^{-\beta|\tau|} \tag{7.67}$$

where β is a constant characteristic of the decay rate [ref. 10, pages 81-82]. Consequently,

$$\psi \equiv e^{-\beta|\tau|} \tag{7.68}$$

By forming the logarithm of (7.67), a first order linear equation is obtained,

$$\ln[\lambda_\varepsilon(\tau)] = -\beta|\tau| + \ln(\sigma_\varepsilon^2) \tag{7.69}$$

which can be used to perform a simple, linear, least squares fit on the empirically derived autocorrelation functions presented earlier. The required algorithm was implemented as a modification to PUSID and performed the fit using the first ten autocorrelation values. β is given by [6],

$$\beta = \left[\frac{\langle \tau \ln[\lambda_\varepsilon(\tau)] \rangle - \langle\tau\rangle\langle \ln[\lambda_\varepsilon(\tau)] \rangle}{\langle\tau^2\rangle - \langle\tau\rangle^2} \right] \tag{7.70}$$

where $\langle\rangle$ denotes an average and τ is a multiple of the interval between successive measurements, z_k. Also, the standard deviation of β then becomes [6],

$$\sigma(\beta) = \frac{n \, \sigma n\{\ln[\lambda_\varepsilon(\tau)]\}}{\{(n-2)[n \Sigma \tau^2 - (\Sigma\tau)^2]\}^{\frac{1}{2}}} \tag{7.71}$$

where $n = 10$. Table 7.2 presents the values of β estimated in this way, together with their associated errors, for all the data files processed using PUSID. Values are included for the log powers in both channels, and the log ratios.

Both of the filter stages described above, including the Mehra algorithm and whiteness test, were incorporated into a single Fortran program, P1516.FOR, originally encoded by Dr. B. J. Rye. A subsequent version, P1516C.FOR, contained several modifications introduced during a series of test runs on both real and simulated data. This latter version was used to process the following data.

7.3.2 Log Power Data

Each run of P1516C generated a series of graphical outputs, characterising the performance of both filtering algorithms; the first example is illustrated in Figures 7.12(a) to 7.12(j) for the real data file, SIFT5.026. Both lidar channels were tuned to the same line (10 R18) for this particular measurement. Figures 7.12(a) and 7.12(b) compare the estimates obtained using the first filter which is limited to processing data with white measurement noise only (and will be referred to hereafter as "P15"). This filter sampled one in every twenty values from the original data sequence to eliminate sequential correlation in the measurement noise (revealed by the earlier plot of its autocorrelation function). Since all original data files contain 10,000 values, sampling at this rate reduces the number of samples to 500.

As with the results obtained using the sub-optimal methods of Section 7.2.2, both estimates are highly correlated, but not identical, revealing a small disparity in the absolute values of the estimates, suggested earlier to be of systematic origin. A similar disparity is also evident in the estimate of Figures 7.12(c) and 7.12(d), generated by the Bryson and Hendrikson algorithm using all 10,000 values in the original data set. This filter is the version capable of accepting coloured measurement noise (and is referred to hereafter as "P16").

An important indication of filter performance is provided by Figures 7.12(e) and 7.12(f) which display an estimate of the variance, p, produced by both P15 (solid line) and P16 (broken line) for each lidar channel. p, therefore, is the "estimate of the variance in the estimate" of the scalar state variable, x, which, in this case corresponds to the log transformed (normalised) power received, in each lidar channel. A further log transformation (base 10) has had to be applied to the ordinate axis, in these and subsequent plots, to compress the range of variation in p occurring between filter initialisation and the steady state condition (beyond the first 2000 samples). However, values corresponding to the steady state square root of p (equivalent to the standard deviation in the state estimate) have also been included.

For both filtering algorithms, P15 and P16, p is dependent on the system and measurement models adopted. The general form of this dependency is apparent from the matrix-vector equations presented in Section 6.3. When reduced to their specific scalar forms for the random walk signal model, this dependency is described by the error covariance extrapolation and update equations, (7.36) and (7.38) for P15, and corresponding equations, (7.62) and (7.63), for P16.

An equivalent set of relationships (7.39) and (7.61) govern the behaviour of the Kalman gain, K, examples of which are shown in Figures 7.12(g) and 7.12(b), for the two lidar channels, 'A' and 'B'. Once again, P15 and P16 outputs are depicted as solid and broken lines, respectively, and log (base 10) compression of the ordinate axis has had to be employed (but steady state values of K are also printed in these figures).

Figures 7.12(e) through to 7.12(h) clearly demonstrate how the two estimates produced by P15 and P16 differ. For this particular example,

both the variance estimate, p, and the gain, K, are less for P16 than for P15. Using the variance estimate as the principal criteria for evaluating filter performance, P16 would therefore appear to produce the better estimate. Furthermore, these estimates appear consistent between lidar channels employing the same wavelength; for example, the steady state value of $\sqrt{p}$ in Figures 7.12(e) and 7.12(f) are in agreement, if allowance is made for error in the estimate of p. Similar comments apply to the gains illustrated in Figures 7.12(g) and 7.12(h).

Both filters are optimised once the innovation sequence of P15,

$$\nu_k = z_k - \hat{x}_{k-1}(+) \tag{7.72}$$

(see (7.37), and the innovation sequence of P16

$$\nu_k = \zeta_{k-1} - h^r \hat{x}_{k-1/k-1}(+) \tag{7.73}$$

(see (7.64)) become white; in other words the presence of any temporal correlation indicates that there is still "information" left in ν [ref. 10, page 317]. Mehra's test for whiteness, introduced in Section 6.3.3 and re-stated in Section 7.3.1, may now be applied to evaluate the significance of any remaining correlation.

Figures 7.12(i) and 7.12(j) present the autocorrelation functions of the innovation sequences of both P15 and P16 for channel A and channel B data respectively. The 95% confidence limits for P15 are depicted by two solid horizontal lines (the upper line coincides with the figure boundary), and those for P16 by two broken lines. The difference in width of these confidence limits is due to, N, the number of data values used to form the autocorrelation functions (this particular application of P1516C used N = 400 for P15 and N = 9000 for P16).

Table 7.3 summarises the important results for each application of

both filters, for log power and log ratio data (discussed in the next session). Ineg is the number of autocorrelation terms which are less than zero and ideally this should equal 10. According to the whiteness tests, therefore, Figures 7.12(a) to 7.12(j) represent a reasonably successful application of both filtering algorithms to the log power data of file SIFT5.026.

Applying P1516C to another real data file, SIFT5.036, obtained using a pair of laser lines recommended for CO_2 concentration measurements (10 R16 - measurement, 10 R8 - reference), the results obtained are as illustrated in Figures 7.13(a) to 7.13(l). For this, and all data files other than SIFT5.026, the first filtering algorithm, P15, sampled one in every ten of the original data set instead of one in twenty (as used for SIFT5.026).

Using the same method of presentation as adopted for Figure 7.12, plots (a) and (b) show the estimate generated by P15 for both lidar channels ((a) corresponds to the measurement channel); plots (c) and (d) show the equivalent results for P16. Consistent positioning and scaling of both axes has been used, where possible, to permit comparisons between superimposed estimates of P15 and P16. Figures 7.13(e) to 7.13(h) illustrate the state error covariance estimates and the Kalman gains for the log powers in each channel. Both sets of estimates reach their steady state values much sooner than those illustrated in Figure 7.12 and these values are larger than those of the previous data file.

As before, an indication of filter performance is provided by the innovation sequence autocorrelation functions of P15 and P16 for both channel A and channel B data (Figures 7.13(i) and 7.13(j)). The Iacf and Ineg values are listed in Table 7.3. According to their Iacf and Ineg counts, P15 performs satisfactorily in both cases, but P16 deviates

from the whiteness condition by Iacf counts of 3 and 4 for the two respective channels. However, upon examination of the innovation acf plots, it becomes apparent that, apart from a single large deviation at lag = 1, these failures are marginal and are therefore not so significant. In fact, for both data files considered so far, and in all subsequent applications of P1516C where whiteness test failure occur, it will be observed that they tend to be marginal for most of the Iacf points counted as lying outside the 95% confidence limits.

Comparing the innovation sequence acf for P16 with the example of Figure 7.9(j), it is apparent that the presence of a single, large acf term at lag = 1 is characteristic of an MA(1) process. This would imply that each innovation sequence term, ν_k, may be interpreted as a linear combination of some current noise variable plus at least one other lagged term of the same noise sequence (see (6.117)). However, this feature does not occur in any of the innovation sequences of P15, and is not present in the P16 results for simulated data (see Section 7.3.4). It would therefore appear to be specific to P16 operating on real data only, and not a consequence of transforming the measurement according to expression (7.56). Although it does not detract from the significance of the whiteness test, it may be regarded as revealing the presence of some time series characteristic not accounted for in the approximation provided by first order autogressive models.

The last two figures, 7.13(k) and 7.13(l), display the results obtained when the log-power estimates in each channel are differenced to provide a log ratio. Although this measurement was conducted using laser lines recommended for CO_2 monitoring, Table 7.1 indicates that, for average humidity levels (50% at 15 C) considerable species interference can be expected from H_2O alone. As can be seen, some excursion

into negative log ratios is evident. This could be due to one of the following:

(a) A displacement in the estimate of systematic origin,

(b) Excessive variance remaining in the estimated signal, or

(c) The presence of absorbers with negative differential absorption coefficients (a manifestation of this particular ratio inversion for the channel A and channel B estimates).

This measurement will be considered further in Section 7.3.3.

One further example of log power estimation is provided here; Figures 7.14(a) to 7.14(l) illustrate the results obtained for a water vapour measurement file, SIFT5.044. The presentation of the figures is identical to those of the previous example; plots (a) through to (d) show the estimates obtained for each channel using both P15 and P16. A notable feature of the error covariance plot, (e), is the close proximity of the variances estimated by P15 and P16. This would imply that both filters offered equivalent performance in terms of the precision of the estimates of the state variable. The estimate of the error covariance is, however, dependent on the accuracies in the estimates of q and r, and errors affecting the performances of both filters, particularly P16, will be discussed later.

Values of Iacf and Ineg, listed in Table 7.3, indicate a relatively poor performance for P16 on the data of channel A but a reasonable result for the data of channel B. Once again, however, if the innovation acf plots (i) and (j) are examined, the Iacf counts appear to be marginal cases. Finally, Figure 7.14(k) and 7.14(l) show the ratios obtained using the difference between the log power estimates in each channel. The initial excursion of the P15 estimates into negative log ratio values

is attributable to the filter adjusting its gain level prior to reaching the steady state condition.

7.3.3 Log ratio Data

The identification of log transformed lidar data with the state variable, x, instead of non-transformed data, was necessary to convert a multiplicative measurement noise process into an additive one, suitable for the implementation of the linear, discrete Kalman filter. A further advantage of this transformation, however, is that the state variable becomes linearly related to either the absorbance A, (expression (3.22(b)) for log power data or the differential absorbance, ΔA (expression (3.24)) for log ratio data. Having, therefore, applied the random walk model to absorptive processes in the previous section, the same model is now applied to differential absorption, since both processes can be expected to exhibit random trend behaviour.

Some of the data files to be examined here have been analysed in previous sections. All are listed in Table 7.1 together with the wavelengths used, their absorption, and differential absorption coefficients. Seven of these data files can be interpreted as gas concentration measurements since they were obtained using recommended [2] measurement/reference line pairs for the specific gas indicated. It must be pointed out, however, that although Petheram is used as a source for these recommended lines, he in fact discusses at length the considerable problems of species interference, particularly due to H_2O, which limit the application of these and other recommended wavelengths, accessible using line tunable CO_2 lasers.

As before, the first application of P1516C examined here is for file SIFT5.026 which was obtained with both channels tuned to the same

laser line (10 R18). The log ratio estimates obtained using P15 and P16 are illustrated in Figures 7.15(a) and 7.15(b), respectively. Since the log ratio for both channels tuned to the same wavelength should obviously be zero, these estimates provide some indication of the magnitude of systematic error to be expected in subsequent estimates. As with previous examples, the estimate of P15 has a larger variance than the estimate of P16, but it is apparent, from both estimates, that some disparity exists between the two channels which, for this particular measurement, results in an erroneous log ratio value of approximately 0.1. A similar value was obtained for another data file, SIFT5.024 (not illustrated), which again used the line (10 R18) in both channels. A comparable log ratio error may therefore be expected in the estimates derived from other files.

Figure 7.15(c) illustrates the lowest steady state error covariances obtained for any of the data files analysed here using P1516C. Although the error covariance estimate is, itself, subject to estimation error, the low values are attributable, partially at least, to speckle correlation between the two channels. Expressions of the form given by (5.22) indicate that positive cross correlation between measurement processes, originating in separate channels, reduces the sum of their individual variance contributions when the measurements are combined as a ratio to form a single estimate. Negative correlation, on the other hand, will increase it. This is in direct contrast to the single channel situation for autocorrelation, discussed previously, where positive correlation reduces sample independence and thereby diminishes the effectiveness of any estimation algorithm operating serially on the data.

The low error covariance estimates of Figure 7.15(c) are therefore a consequence of the correlation reduced estimates of the speckle

measurement noise, r. Since the Kalman gains are inversely proportional to this measurement noise they also achieve their lowest steady state values as revealed by Table 7.3 and illustrated in Figure 7.15(d).

The innovation acf of P15 (Figure 7.15(e)) yielded an Iacf count of zero but a value of Ineg = 15, indicating the presence of some bias. For P16, however, a reasonable result was obtained with Iac = 2 and Ineg = 10.

The data file SIFT5.036 was examined previously by forming separate estimates of the log powers in each channel and then differencing these estimates to produce the Figures 7.13(k) using P15, and 7.13(l) using P16. In this section the log ratio is estimated directly and the results are as presented in Figures 7.16(a) and (b). Both of these ratio estimates possess larger variances than their equivalents in Figure 7.13 due, presumably, to the higher measurement noises associated with log ratio data. Values for the square root of r, listed in Table 7.3, have averages of 0.53 for log power data and 0.70 for log ratio data. In terms of the steady state error covariances alone, therefore, the log ratio estimates would appear to be inferior to the alternatives formed by differencing the estimates of the log powers. Other factors, however, such as the bias associated with each of the estimation and ratioing methods would also have to be taken into consideration.

Figure 7.16(c) confirms these relatively high variances but the gain histories illustrated in Figure 7.16(d) have values falling midway between the two sets of gain estimates illustrated in Figures 7.13(g) and (h). In terms of filter performance P15 satisfies the requirements of the innovation sequence whiteness test (Iacf = 0, Ineg = 10) and, with the single exception of the deviation at lag = 1, P16 almost does as well (Iacf = 1, Ineg = 10). The magnitude of the autocorrelation terms is illustrated in Figure 7.16(e).

In addition to providing a reference for filenames, laser lines and absorption coefficients, Table 7.1 also lists example values of the log ratio expected for "typical" concentrations of the four gases considered here: H_2O, CO_2, NH_3 and C_2H_4. The typical value assumed for water vapour concentration corresponds to a relative humidity of 50% at 15°C. Ambient concentrations of CO_2 and NH_3 were extracted from reference [12], but the value selected for C_2H_4 was equated to that for NH_3 because no alternative estimates were available at the time of compiling the table. It is appreciated that an ethylene concentration of $\sim$2 ppb is probably an over-estimate as a global ambient figure. However, these measurements were conducted in close proximity to various industrial sites so the local concentrations may have been higher.

Using expression (3.32), the log ratio due to n interfering species is given by

$$\text{log ratio} = 2[\gamma_1\Delta\alpha_1' + \gamma_2\Delta\alpha_2' + \cdots + \gamma_n\Delta\alpha_n']r \qquad (7.74)$$

Although the measurement data of file SIFT5.036 was obtained using laser lines recommended for CO_2 monitoring, Table 7.1 reveals that considerable interference can be expected from water vapour. In fact, for the water vapour concentration assumed in the table, both CO_2 and H_2O have comparable log ratios. No attempt is made, therefore, to convert the log ratio estimates into concentration estimates. However, summing the log ratio contribution from all four gases yields a value close to 0.2. The estimates of Figures 7.16(a) and 7.16(b) appear to fluctuate around a mean value close to 0.2, occasionally descending into negative log ratio values. Possible factors contributing to these excursions into the negative log ratio region were itemised in Section 7.3.2 but some component of the fluctuations observed in these estimates could well be

due to the irregular distribution of H_2O, or some other gas of industrial origin, over the period of the measurement. The highly correlated results obtained for files such as SIFT5.026 and SIFT5.024, which used identical wavelengths in each channel, at least suggest that these fluctuations are not of systematic origin.

Similar comments apply to the P1516C results for the data files SIFT5.037 and SIFT5.038, illustrated respectively in the figure sequences 7.17 and 7.18. SIFT5.037 used laser lines recommended for the measurement of ammonia (NH_3) concentration; log ratio estimates are presented in Figures 7.17(a) and (b) and the filter performance characteristics, P_∞, K_∞ and innovation sequence acf) in Figures 7.17(c) to 7.17(e). Once again, reasonable performance is observed for both filters. The first estimate, of filter P15, shows a rapid descent into negative log ratio values followed by an equally rapid recovery as the filter gain settles down to its steady state value.

If Table 7.1 is now examined for the data file SIFT5.037 it becomes apparent that the major contribution to the log ratio is due to H_2O, with CO_2 providing the second largest contribution. The log ratio component due to NH_3 is, in fact, negative and less than 10% of that due to H_2O (for the concentration values assumed in the table).

Log ratio estimates for the data file SIFT5.038 are illustrated in Figures 7.18(a) and 7.18(b). Here, also, negative log ratio values occur, the magnitude of which depends on which estimate, P15 or P16, is considered. Clearly for these estimates, and those of SIFT5.036 (Figure 7.16(a) and (b)), excessive noise remains, (particularly for P15). As before, this is reflected in the error covariance estimates of Figures 7.18(c) which are somewhat larger than those observed in other log ratio estimates (see Table 7.3 - some of these have yet to be discussed).

Figure 7.18(d) presents the Kalman gains for P15 and P16. Although the Iacf number for P16 (Table 7.3) indicates that the innovation sequence was not as white as it could be, examination of Figure 7.18(e) reveals that the result is comparable to the performance of P16 for other data files.

Example log ratio values listed in Table 7.1, once again, indicate that the contributions due to CO_2 and H_2O can be expected to dominate. Forming the sum of these log ratios, due to each of the four gases, suggests that the estimates of Figures 7.18(a) and (b) should have values of the order of $\sim$0.2 and, allowing for fluctuations, this seems to be confirmed.

Figure sequences 7.19, 7.20, 7.21 and 7.22 present the results of applying P1516C to four water vapour measurement files, SIFT5.043, 044, 045, and 046. The first three used the wavelength pairs 10 R20 (measurement) and 10 R18 (reference) but, for SIFT5.046, the reference wavelength was shifted to 10 R16. As can be seen from Table 7.1, relatively little species interference can be expected from the other three gases. Log ratio estimates for P15 and P16 are presented in plots (a) and (b) for each of these figures.

In each case the "noisiness" of the state variable estimate is reflected in the error covariance estimates (p). Steady state values of p are listed in Table 7.3 for both filters but their complete histories during each filter run are illustrated in plots '(c)' for Figures 7.19 to 7.22. A spread of steady state error covariances are represented with values of $\sqrt{p_\infty}$ ranging from 4.42×10^{-2} for SIFT5.043 to 0.1783 for SIFT5.046. It will be noticed that, as $\sqrt{p_\infty}$ decreases, both filters take relatively longer to settle into the steady state condition. In Figure 7.22(c), for example, $\sqrt{p_\infty}$ appears to settle after approximately 500

samples but, for 7.19(c), the steady state condition is not reached until beyond a sample index of k = 3000. Similar observations apply to the Kalman gains illustrated in all plots '(d)' for Figures 7.19 to 7.22.

Filter performances were, in general, not completely optimised according to the whiteness test counts listed in Table 7.3 but the innovation sequence acfs (plots '(e)' in Figures 7.19 to 7.22) reveal only slight deviations from the 95% confidence limits. In one instance (SIFT5.046) the performance of P16 appears superior to that of P15 if the innovation sequence acf, alone, is considered.

Concentration estimates derived from P15 and P16 log ratio estimates are presented in plots (f) and (g) of Figures 7.19 through to 7.22. As with previous data files, estimates formed before the filter has reached its steady state may be ignored. These concentrations are further interpreted as relative humidity profiles in plots '(h)' and '(i)' of the same figures using the expression

$$RH(\%) = \gamma \frac{AP}{SVP(T)} 100 \qquad (7.75)$$

where RH is relative humidity (expressed as a percentage), γ is the concentration of water vapour, AP is atmospheric pressure (= 0.101325 MPa for the International Civil Aviation Organisation standard atmosphere) and SVP(T) is the saturated vapour pressure at temperature, T(^{o}C).

A hair-based hygrometer was available at the lidar site (situated outside the lidar laboratory window), capable of monitoring both relative humidity and temperature using a recording drum set to rotate once in 24 hours. Listed below are the relative humidities and temperatures recorded during each of the measurements indicated, together with the concentration and log ratio values they correspond to:

Measurement file number	RH(%)	T(^{o}C)	γ	Log ratio
043	72	4	5.78×10^{-3}	1.57
044	68	5	5.85×10^{-3}	1.59
045	75	9.5	8.80×10^{-3}	2.39
046	72	10	8.72×10^{-3}	2.26

As can be seen, although the filtered log ratios yield reasonable relative humidities, a significant difference exists between these estimates and the values recorded by the hygrometer (which was accurate to within a few per cent).

It is possible that these differences are a consequence of attempting to compare the integrated absorption over a path length of 3.6 km with a point sensor positioned at one extremity of this absorption path. Systematic log ratio errors of $\sim \pm 0.1$, however, will also contribute a maximum error of ±10% in both the concentration and relative humidity estimates. Bias, resulting from the use of a log ratio estimator, is another potential source of error.

Fluctuations, apparent in each of the estimates, vary from measurement to measurement but are closely related to the relative magnitudes of the steady state error covariances. If the minimum variance case is considered (SIFT5.043), it would appear that some non-stationarity is present which produces a variation in the relative humidity of approximately 10% during the latter half of the experiment. Some variation was also evident in the hygrometer reading over an equivalent period of time (≈1%) but since this device appeared to have a characteristic time response of at least 15 minutes (observed during the onset of heavy rain), it is difficult also to compare transient behaviour between the two types of measurement system.

7.3.4 Simulated Log Power Data

When analysing real data, the system and measurement noise variances (q and r) are, of course, unknown so it is difficult to determine how accurate the Mehra estimates of these quantities are and, consequently, the accuracy of the error covariance estimate, p_∞. Some insight, however, may be obtained from the use of simulated data in which system and measurement models are completely specified.

The two simulated log power files used here have already been considered in the analysis of previous sections. Both use the same random walk signal model (illustrated in Figure 7.7(a)) but one is additively combined with uncorrelated (white) speckle measurement noise and the other with correlated (coloured) speckle measurement noise. They will be referred to here, respectively, as SIM.001 and SIM.002. The system noise model used a driving noise variance of $q = 10^{-5}$ and the measurement noise sequence had a variance, $r = 0.25$ so that $\sqrt{q} = 3.16 \times 10^{-3}$ and $\sqrt{r} = 0.5$. For the correlated noise version, the correlation coefficient was specified to decay to 0.1 after five "shots" so that, using expression (7.68), $\psi = 0.631$.

P15 estimates of the two simulated files are illustrated in Figures 7.23(a) and (b), with the latter corresponding to coloured measurement noise. The equivalent P16 estimates are available in the following plots 7.23(c) and (d). As noted in Section 7.2.2, those plots for the correlated measurement noise cases show a noisier estimate than the uncorrelated versions. The P16 estimates are re-plotted in Figures 7.23(e) and 7.23(f) to compare the estimates (solid line) with the true signal (broken lines).

Of primary importance, however, are the error covariance plots of Figures 7.23(g) (uncorrelated noise) and 7.23(h) (correlated noise).

For both P15 and P16, the steady state error covariances are worse for the case in which correlated measurement noise is present but are comparable to the values obtained for real data. It will be noticed from Table 7.3 that the Mehra estimates for $\sqrt{q}$ are 4.75×10^{-3} for SIM.001 and 7.91×10^{-3} for SIM.002 (these P16 values have been scaled down from the P15 estimates using the relationship (7.52)). The first value, obtained in the presence of uncorrelated measurement noise, is approximately 50% larger than the actual value of $\sqrt{q}$ used, whereas the value derived from the correlated measurement noise file is nearly 250% larger! Autocorrelation values occurring over the first 10 lags for SIM.002 (Figure 7.11(j)) are, however, significantly higher than for many of the real log power estimates.

These two extremes of error, therefore, may be used as some indication of the possible upper and lower limits in the sense that estimates of q generated by the Mehra algorithm tend, here, to be larger than the real values. As with real data, the smaller the estimate of q, the smaller the error covariance and the Kalman gains (Figures 7.23(i) and 7.23(j)), resulting in less noise in the estimates.

Filter performance characteristics were slightly different from those obtained with real data; although P15 performed reasonably well for both files, the P16 estimates appear to possess biased innovation sequences for both the uncorrelated and correlated version. However, whereas the Iacf value for SIM.001 is 0, that for SIM.002 is 3. The innovation sequence autocorrelation functions displayed in Figures 7.23 (k) and 7.23(l), reveal that although the P15 results for both simulated files are very similar to those of the real data, the P16 acfs lack the large negative value at lag = 1. SIM.002 does, however, possess a small, positive acf delay over approximately the first four lags, characteristic of an autoregressive process.

The lasts two plots, (m) and (n), of Figure 7.23 compare the error in the estimate of P16 with the error covariance estimate $\sqrt{p}$ for both files, SIM.001 and SIM.002. The quantity expressed by the ordinate axis is

$$\frac{\hat{x}_k - x_k}{\sqrt{p_\infty}} \tag{7.76}$$

where $\hat{x}_k$ is the state variable estimate and x_k is the true random walk signal. Despite differences in filter performance the actual estimate errors are very similar for both versions of the simulated data, and would also appear to have rms values approximately equivalent to the estimated error covariance.

7.3.5 Concluding Remarks on the Optimum Estimation Techniques Used

The Mehra algorithm used with P15 provides an estimate of the variances of both the system and measurement noise processes, q and r, which directly influence the magnitude of the error covariance, p. This error covariance is important because it provides an estimate of the error in the state variable (log power or log ratio) which can then be translated into an estimate of the precision of any gas concentration value derived from the state variable.

It was observed in the previous section, 7.3.4, that errors in the Mehra estimates of q cause $\hat{q}$, and the estimates of both p_∞ and K_∞, to be larger than they should be for an optimised filter. Such errors result in excessive measurement noise remaining in the estimate of the state variable. The Mehra algorithm is intended to be used in an iterative scheme in which the estimates of q and r asymptotically approach

their optimum values. In principle, to move along the asymptote, it is necessary to re-run the filter (in this case, P15), each time using the latest recommendations for q and r, until the innovation sequence is white.

Figures 7.24(a) and (b) present two examples of how the estimates of q improve with each iteration, obtained using two real data files. It was found that although these estimates did converge in general, they very rarely converged to a value which resulted in a white innovation sequence. In order to make the transition from the convergence value to the "optimised" value it was necessary to manually tune the system by applying certain rules of thumb [5]. These, in essence, use the bias in the innovation sequence (Ineg) to indicate whether the system noise recommendation should be increased or decreased. Assuming the measurement noise is approximately correct, negative bias in the innovation sequence means that $\hat{q}$ should be reduced (and vice versa). Optimised values of q obtained by such means are indicated in both figures by the dashed lines.

Estimates of $\sqrt{q}$ in Table 7.3 vary by an order of magnitude; some of this variation will be attributable to errors in the estimate but the rest may be characteristic of variations in the system process itself. Mehra estimates of the values of r converge quickly and appear reasonably consistent, yielding approximate values of 0.5 for log power data, and 0.7 for log ratio data.

Using the expression (7.5), values of $\sqrt{p_\infty}$ can be interpreted directly as percentage errors in the "absorbance" A, or "differential absorbance", ΔA, of equations (3.22(b)) and (3.24) since,

$$\sqrt{p_\infty} = \sigma(\text{log power}) = \frac{\sigma(A)}{\bar{A}} \qquad (7.77)$$

$$\text{or} \quad \sqrt{p_\infty} = \sigma(\log \text{ ratio}) = \frac{\sigma(\Delta A)}{\bar{\Delta}\bar{A}} \tag{7.78}$$

where $\bar{A}$ and $\bar{\Delta}\bar{A}$ denote time dependent expectation values of the signal at any given instant. Therefore, using the P16 estimate of $\sqrt{p_\infty}$ (Table 7.3), these errors would appear to vary from nearly ±18% for SIFT5.046 down to ±3.6% for SIFT5.026 (due to the fact that, for this file, the measurement noise in both channels is highly correlated).

Steady state values for p and K can be predicted for the linear Kalman filter using the matrix Riccati equation which, in scalar form, is given by expression (7.40). In Section 7.3.1 it was shown that, for the random walk system model,

$$p_\infty = \sqrt{qr} \tag{7.43}$$

and that the scaling relationship between the estimate, $\hat{q} = q_n$ derived from P15, and the value q_1 used by P16, is

$$q_1 = \frac{q_n}{n} \tag{7.52}$$

where n is the sample interval used on the original data set by P15.

Bryson and Hendrikson do not consider the continuous filter case and therefore do not provide an expression for the continuous propagation of covariance. However, since the linear Kalman filter represents the optimum filter [ref. 10, p. 107], the Bryson and Hendrikson algorithm will not do better than the result, (7.43). Therefore, with P15 and P16 both performing as optimum linear filters, the ratio between the two estimates of p will achieve a maximum value of

$$\frac{p_\infty \text{ (P16)}}{p_\infty \text{ (P15)}} = \frac{\sqrt{\left(\frac{q_n}{n}\right) r}}{\sqrt{q_n r}} = \frac{1}{\sqrt{n}}$$

which, for n = 10, is ≈0.32. For the purpose of comparing estimates

listed in Table 7.3, the square root of this ratio (= 0.56 for n = 10) is more useful. The results tabulated tend to confirm that P16 does, in fact, provide a better estimate than P15 since the error covariance estimates generated by P16 are consistently lower than those of P15. With a single exception, however, ratios between these estimates do not achieve the theoretical maximum of $\sim$0.56. The one exception is provided by the simulated data file, SIM.001, for which the speckle sequence is uncorrelated.

Sequential correlation in speckle measurement noise does, therefore, detract from the information content of each measurement event, z_i, even if specialised algorithms are used which are designed to cope with it. It is equally apparent, however, that such algorithms help minimise the influence of this noise imposed sample dependency and therefore permit useful sampling within the correlation time constant (= $1/\beta$ - see expression (7.68)). P16 is, however, subject to at least one more source of error than P15 since estimates of β (the speckle decorrelation parameter) must be provided. These are listed in Table 7.2 together with their error estimates. In adopting the random walk model for both P15 and P16, the autoregressive parameter ϕ in (7.25) is assumed to be unity. If, for any real data file, a better system model exists using a value of $\phi \neq 1$, then any errors arising from the use of the random walk model ($\phi = 1$) will be common to both filters.

It is worth noting here, however, that since the general autoregressive model (7.43) with $\phi < 1$ has the equivalent continuous form

$$\dot{x} = -\alpha x + w \tag{7.79}$$

for which $\alpha > 0$ (this is derived from (7.41) with $f = -\alpha$ and $g = 1$), solving the matrix Riccati equation for the steady state condition yields

$$p_\infty = -\alpha r \pm \sqrt{\alpha^2 r^2 + qr} \tag{7.80}$$

Therefore, as this system model departs from the random walk condition ($\alpha = 0$) and tends towards the white noise sequence, w ($\alpha = 1$), the steady state error covariance decreases, thereby potentially improving the precision of the estimate. Figure 7.25 is a plot of the dependency of p_∞ as a function α for the values of q and r used in the random walk simulations.

Finally, taking the expression (3.32) which relates the concentration of a gas species, m, to the log ratio, and applying a standard result for the component variances in a linear expression [6], the precision in the concentration measurement can be related to the steady state error covariance via

$$\sigma^2(\gamma_m) = \frac{p_\infty}{(2\Delta\alpha' r)^2} \tag{7.81}$$

where $\Delta\alpha'$ is the differential absorption coefficient per cm.

This function is plotted in Figure 7.26 (as a standard deviation rather than a variance) for each of the four gases identified in Table 7.1 ($\Delta\alpha' = 7.56 \times 10^{-4}$ for H_2O) using values of $\sqrt{p_\infty}$ ranging between 1 and 10^{-4}. A log transformation has once again been applied to the axes. Since the differential absorption coefficients for H_2O and CO_2 are nearly identical for the measurement wavelengths chosen, both plots appear coincident. Shaded areas translate the upper and lower bounds for $\sqrt{p_\infty}$ (P16), as listed in Table 7.3, into precision bounds for all four gases. These precision estimates assume, of course, that p_∞ completely defines all sources of error.

REFERENCES

[1] RYE, B.J. and THOMAS, E.L. "Application of coherent lidar to differential absorption" published in 'Optical and Laser Remote Sensing', in the series Topics in Applied Physics, Springer-Verlag, 1982.

[2] PETHERAM, J.C. "A comparison between continuously tunable and line tunable CO_2 laser systems for the remote sensing of atmospheric trace gases". Report prepared at Hull University (Department of Applied Physics), England, for the National Physical Laboratory, Teddington, England, 1981.

[3] BOSCHER, J. Compilation of tables listing the absorption coefficients of various atmospheric constitutions, pollutants and trace gases at CO_2 laser line tunable wavelengths between 9 and 11 microns. Obtained from the Battelle Institute, Frankfurt and Main,

[4] MENYUK, N., KILLINGER, D.K. and MENYUK, C.R. "Limitations of signal averaging due to temporal correlation in laser remote-sensing measurements", Applied Optics, 21(18), 15 September 1982, pp. 3377-3383.

[5] RYE, B.J. Private communication.

[6] BARFORD, N.C. "Experimental Measurements: Precision Error and Truth", Addison-Wesley, 1967.

[7] BOX, G.E.P. and JENKINS, G.M. "Time Series Analysis: Forecasting and Control", San Francisco: Holden-Day, Inc., 1970.

[8] VANDAELE, W. "Applied Time Series and Box-Jenkins Models", Academic Press, ISBN 0 12 712650 3, 1983.

[9] YOUNG, P. "Recursive Estimation and Time Series Analysis: An Introduction", Springer-Verlag, 1984.

[10] GELB, A. "Applied Optimal Estimation", The MIT Press, ISBN 0 262 70008 5, 1974, p. 105.

[11] MEHRA, R.K. "On the identification of variances and adaptive Kalman filtering", IEEE Transactions on Automatic Control, AC-15(2), April 1970, pp. 175-184.

[12] BRYSON, Jr. A.E. and HENRIKSON, L.J. "Estimation using sampled data containing sequentially correlated noise", J. Spacecraft, 5(6), June 1968, pp. 662-665.

PAGE NUMBERING AS ORIGINAL

TABLE 7.1

File Number	Laser Lines: Channel A / Channel B	Wavelength (microns) Channel A / Channel B	H_2O ($\gamma = 8.4 \times 10^{-3}$ @ 15 C) 50% RH Absorp. Coeff. (cm^{-1})	D.A.C. (cm^{-1})	Log ratio	CO_2 ($\gamma = 0.000318$) Absorp. Coeff. (cm^{-1})	D.A.C. (cm^{-1})	Log ratio	NH_3 ($\gamma = 2 \times 10^{-9}$) Absorp. Coeff. ($cm^{-1}$)	D.A.C. (cm^{-1})	Log ratio	C_2H_4 ($\gamma = 2 \times 10^{-9}$) Absorp. Coeff. ($cm^{-1}$)	D.A.C. (cm^{-1})	Log ratio
036	10 R16 10 R8	10.2744 10.3337	1.31×10^{-4} 1.03×10^{-4}	2.8×10^{-5}	0.085	3.05×10^{-3} 2.30×10^{-3}	7.5×10^{-4}	0.086	0.110 23.20	-23.09	-0.017	1.09 1.05	0.04	2.88×10^{-5}
037	10 R12 10 R8	10.3035 10.3337	1.66×10^{-4} 1.03×10^{-4}	6.3×10^{-5}	0.191	2.87×10^{-3} 2.30×10^{-3}	5.7×10^{-4}	0.065	0.230 23.2	-22.97	-0.016	1.96 1.05	0.95	6.84×10^{-4}
038	10 P14 10 P28	10.5321 10.6746	1.22×10^{-4} 0.94×10^{-4}	2.8×10^{-5}	0.085	2.74×10^{-3} 1.83×10^{-3}	9.1×10^{-4}	0.104	0.870 0.360	0.510	3.67×10^{-4}	36.5 1.34	35.16	0.025
043	10 R20 10 R18	10.2466 10.2604	8.51×10^{-4} 0.95×10^{-4}	7.56×10^{-4}	2.289	2.87×10^{-3} 3.02×10^{-3}	-1.5×10^{-4}	0.172	0.050 0.063	-0.013	-9.36×10^{-6}	1.16 0.70	0.46	3.31×10^{-4}
044	10 R20 10 R18	10.2466 10.2604	8.51×10^{-4} 0.95×10^{-4}	7.56×10^{-4}	2.289	2.87×10^{-3} 3.02×10^{-3}	-1.5×10^{-4}	0.172	0.050 0.063	-0.013	-9.36×10^{-6}	1.16 0.70	0.46	3.31×10^{-4}
045	10 R20 10 R18	10.2466 10.2604	8.51×10^{-4} 0.95×10^{-4}	7.56×10^{-4}	2.289	2.87×10^{-3} 3.02×10^{-3}	-1.5×10^{-4}	0.172	0.050 0.063	-0.013	-9.36×10^{-6}	1.16 0.70	0.46	3.31×10^{-4}
046	10 R20 10 R16	10.2466 10.2744	8.51×10^{-4} 1.31×10^{-4}	7.20×10^{-4}	2.179	2.87×10^{-3} 3.05×10^{-3}	-1.8×10^{-4}	-0.021	0.050 0.110	-0.060	-4.32×10^{-5}	1.16 1.09	0.07	$5.04 \times 10^{0-5}$

D.A.C. = Differential Absorption Coefficient (per cm)

TABLE 7.2

β Estimates

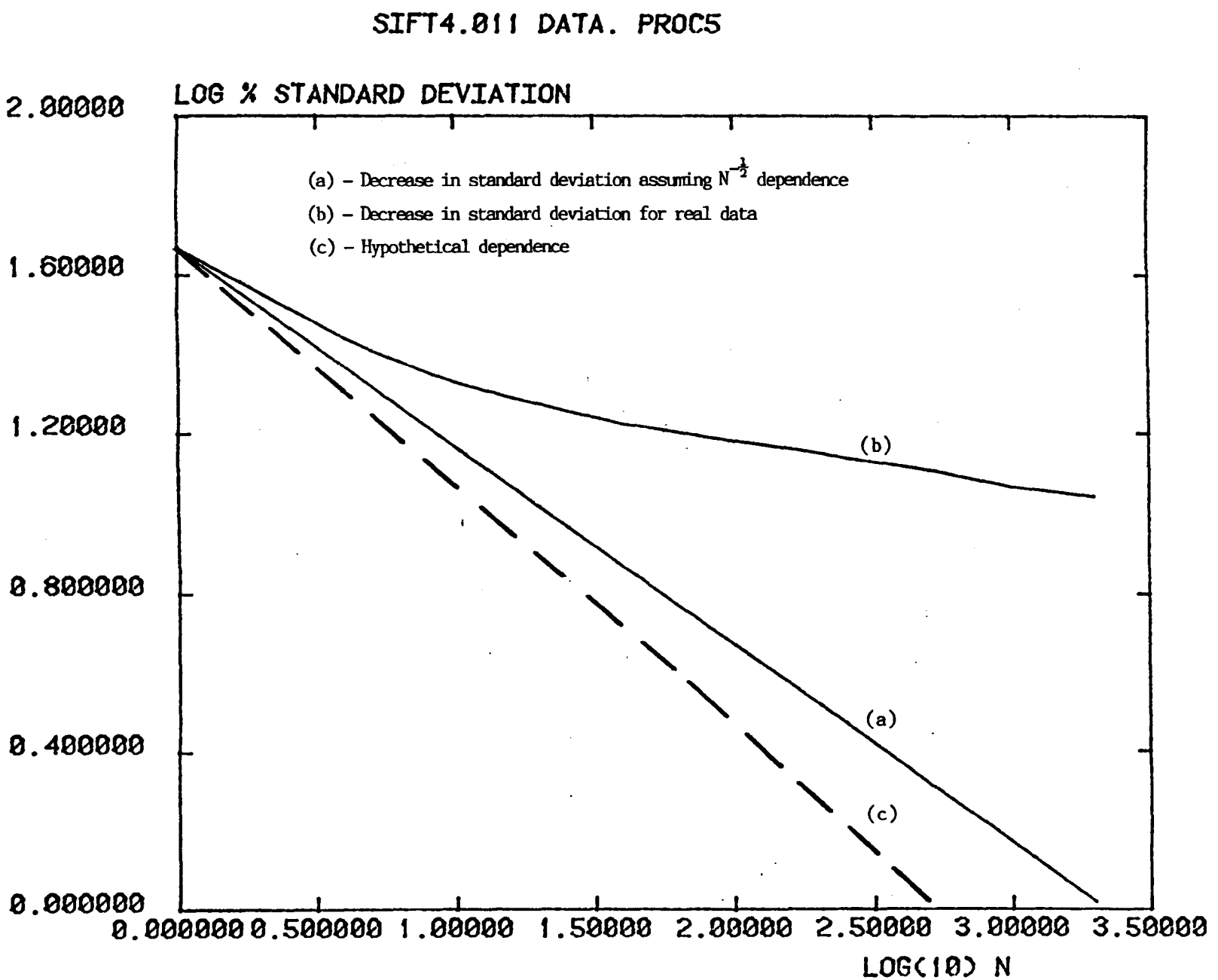
Filename	ln(Pa)	ln(Pb)	ln(ratio)
SIFT5.024	-0.30 ± 0.10	-0.26 ± 0.09	-0.29 ± 0.11
SIFT5.026	-0.20 ± 0.07	-0.20 ± 0.07	-0.27 ± 0.11
SIFT5.036	-0.25 ± 0.10	-0.27 ± 0.11	-0.22 ± 0.10
SIFT5.037	-0.23 ± 0.09	-0.34 ± 0.14	-0.26 ± 0.11
SIFT5.038	-0.22 ± 0.09	-0.26 ± 0.11	-0.21 ± 0.10
SIFT5.043	-0.25 ± 0.10	-0.32 ± 0.12	-0.35 ± 0.15
SIFT5.044	-0.21 ± 0.08	-0.24 ± 0.10	-0.27 ± 0.11
SIFT5.045	-0.26 ± 0.09	-0.26 ± 0.10	-0.29 ± 0.11
SIFT5.046	-0.21 ± 0.08	-0.25 ± 0.10	-0.24 ± 0.09

TABLE 7.3

Summary of Filter Performance

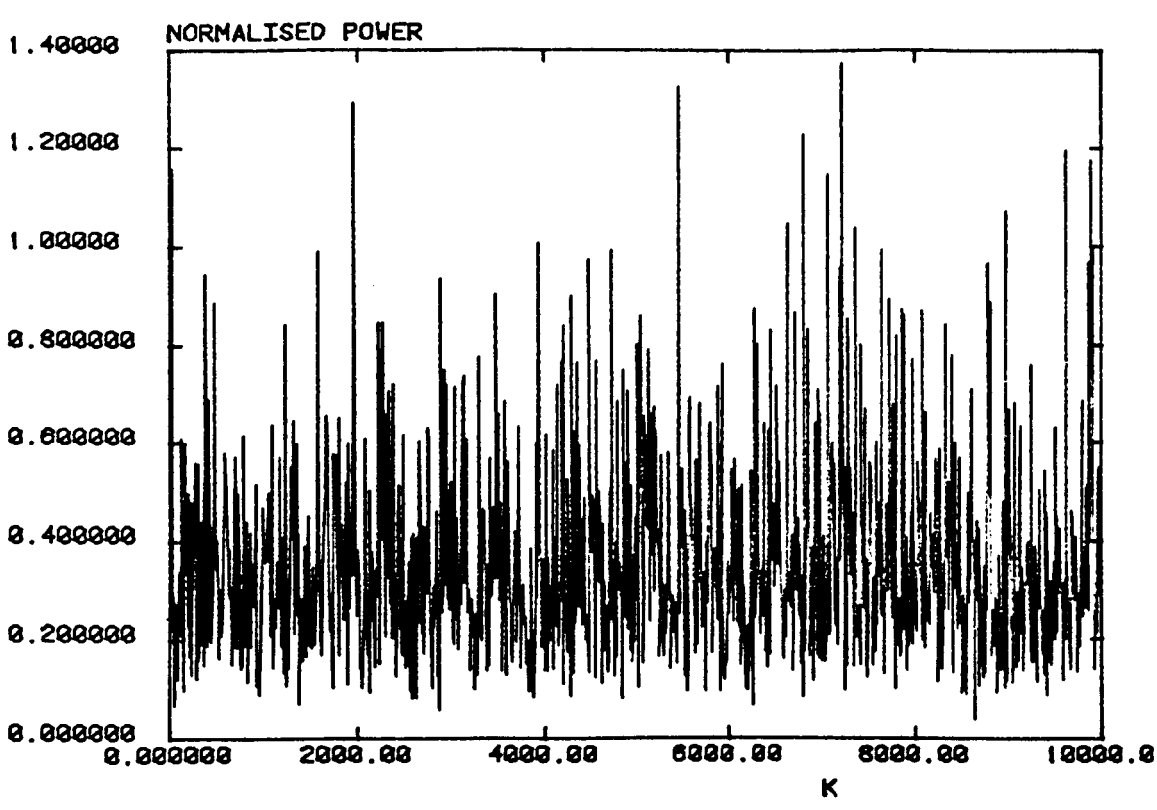
		P15						P16					
Filename	State Variable	$\sqrt{q}$	$\sqrt{r}$	$\sqrt{P_\infty}$	K_∞	I_{acf}	I_{neg}	$\sqrt{q}$	$\sqrt{r}$	$\sqrt{P_\infty}$	K_∞	I_{acf}	I_{neg}
SIFT5.026	ln(Pa)	1.0×10^{-2}	0.49	7.00×10^{-2}	2.04×10^{-2}	0	9<0	2.24×10^{-3}	0.49	5.79×10^{-2}	7.86×10^{-2}	1	9<0
	ln(Pb)	1.0×10^{-2}	0.46	6.82×10^{-2}	2.15×10^{-2}	0	9<0	2.24×10^{-3}	0.46	5.70×10^{-2}	8.40×10^{-2}	1	9<0
SIFT5.035	ln(Pa)	1.63×10^{-2}	0.57	9.68×10^{-2}	2.84×10^{-2}	1	9<0	5.17×10^{-3}	0.57	9.03×10^{-2}	1.43×10^{-2}	3	11<0
	ln(Pb)	3.95×10^{-2}	0.56	0.1507	6.87×10^{-2}	0	11<0	1.25×10^{-2}	0.56	0.1345	3.33×10^{-2}	4	9<0
SIFT5.044	ln(Pa)	1.25×10^{-2}	0.56	8.43×10^{-2}	2.20×10^{-2}	1	12<0	3.95×10^{-3}	0.56	8.25×10^{-2}	1.19×10^{-2}	6	8<0
	ln(Pb)	4.0×10^{-2}	0.53	0.1485	7.25×10^{-2}	1	12<0	1.26×10^{-2}	0.53	0.1367	3.70×10^{-2}	2	11<0
SIFT5.026	log ratio	6.94×10^{-3}	0.22	3.91×10^{-2}	3.15×10^{-2}	0	16<0	2.19×10^{-3}	0.22	3.56×10^{-2}	1.53×10^{-2}	2	10<0
SIFT5.036	log ratio	3.30×10^{-2}	0.76	0.1604	4.23×10^{-2}	0	10<0	1.04×10^{-2}	0.76	0.1533	2.24×10^{-2}	1	10<0
SIFT5.037	log ratio	1.0×10^{-2}	0.72	8.49×10^{-2}	1.39×10^{-2}	0	11<0	3.16×10^{-3}	0.72	7.88×10^{-2}	6.85×10^{-3}	1	8<0
SIFT5.038	log ratio	3.50×10^{-2}	0.74	0.1626	4.63×10^{-2}	0	10<0	1.11×10^{-2}	0.72	0.1560	2.47×10^{-2}	2	10<0
SIFT5.043	log ratio	4.0×10^{-3}	0.65	5.12×10^{-2}	6.13×10^{-3}	1	9<0	1.26×10^{-3}	0.65	4.42×10^{-2}	2.72×10^{-2}	2	9<0
SIFT5.044	log ratio	2.0×10^{-2}	0.72	0.1204	2.76×10^{-2}	0	11<0	6.32×10^{-3}	0.72	0.1103	1.35×10^{-2}	4	9<0
SIFT5.045	log ratio	1.14×10^{-2}	0.70	8.99×10^{-2}	1.62×10^{-2}	0	9<0	3.60×10^{-3}	0.70	8.14×10^{-2}	7.70×10^{-3}	3	12<0
SIFT5.046	log ratio	5.0×10^{-2}	0.72	0.1930	6.71×10^{-2}	3	10<0	1.58×10^{-2}	0.72	0.1783	3.43×10^{-2}	3	10<0
SIM.001	ln(power)	1.50×10^{-2}	0.48	8.60×10^{-2}	3.06×10^{-2}	1	10<0	4.75×10^{-3}	0.48	4.76×10^{-2}	9.67×10^{-3}	0	15<0
SIM.002	ln(power)	2.50×10^{-2}	0.49	0.1117	5.01×10^{-2}	1	11<0	7.91×10^{-3}	0.49	8.21×10^{-2}	1.87×10^{-2}	3	14<0

FIGURE 7.1



Standard Deviation as a function of the Number of Samples Averaged

FIGURE 7.2 (a)



Speckle Induced Normalised Power: in Channel A (above)

: in Channel B (below)

FIGURE 7.2 (b)

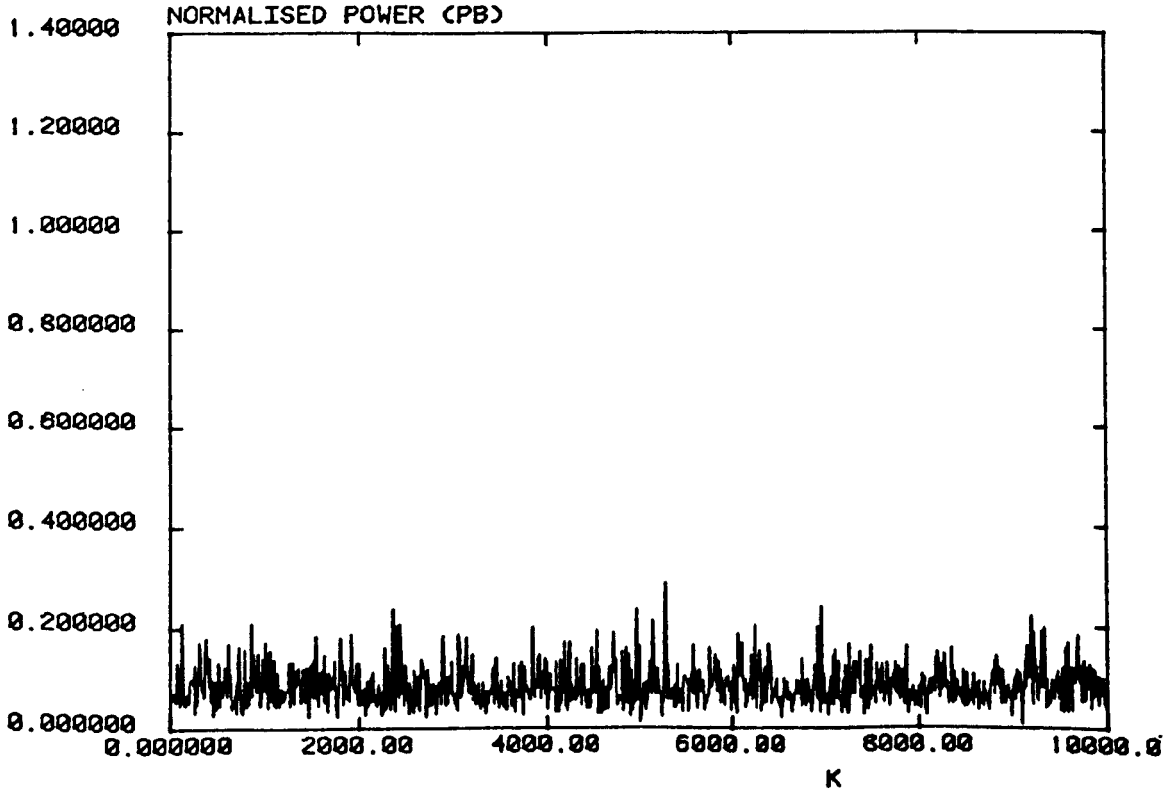
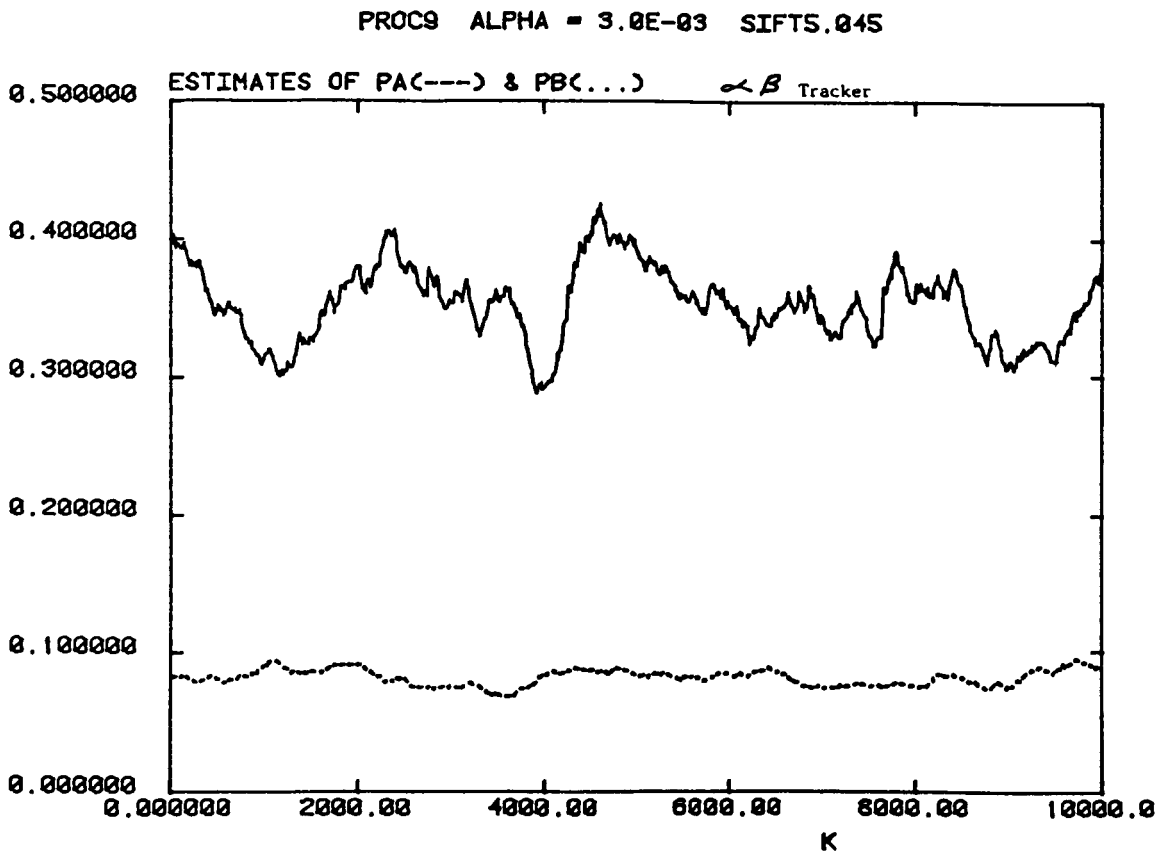


FIGURE 7.2 (c)



Estimates of the Normalised Power: $\propto \beta$ Tracker (above)

: Running Average (below)

FIGURE 7.2 (d)

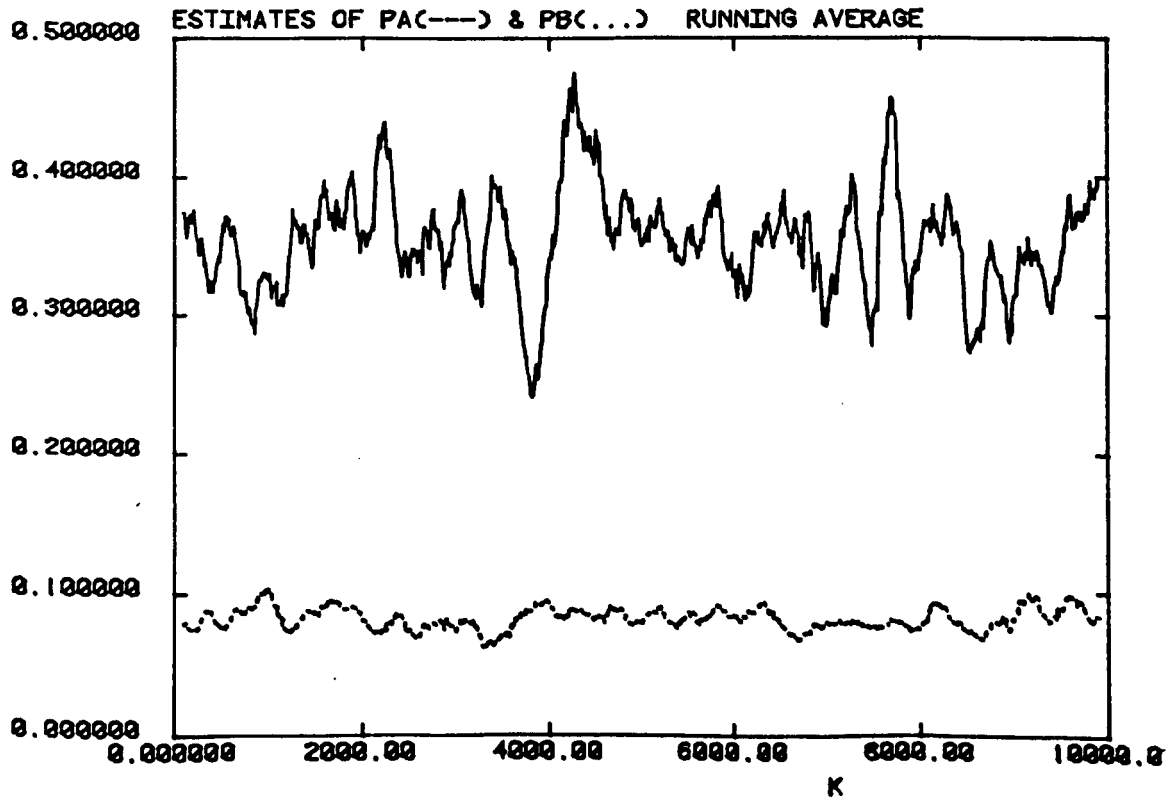
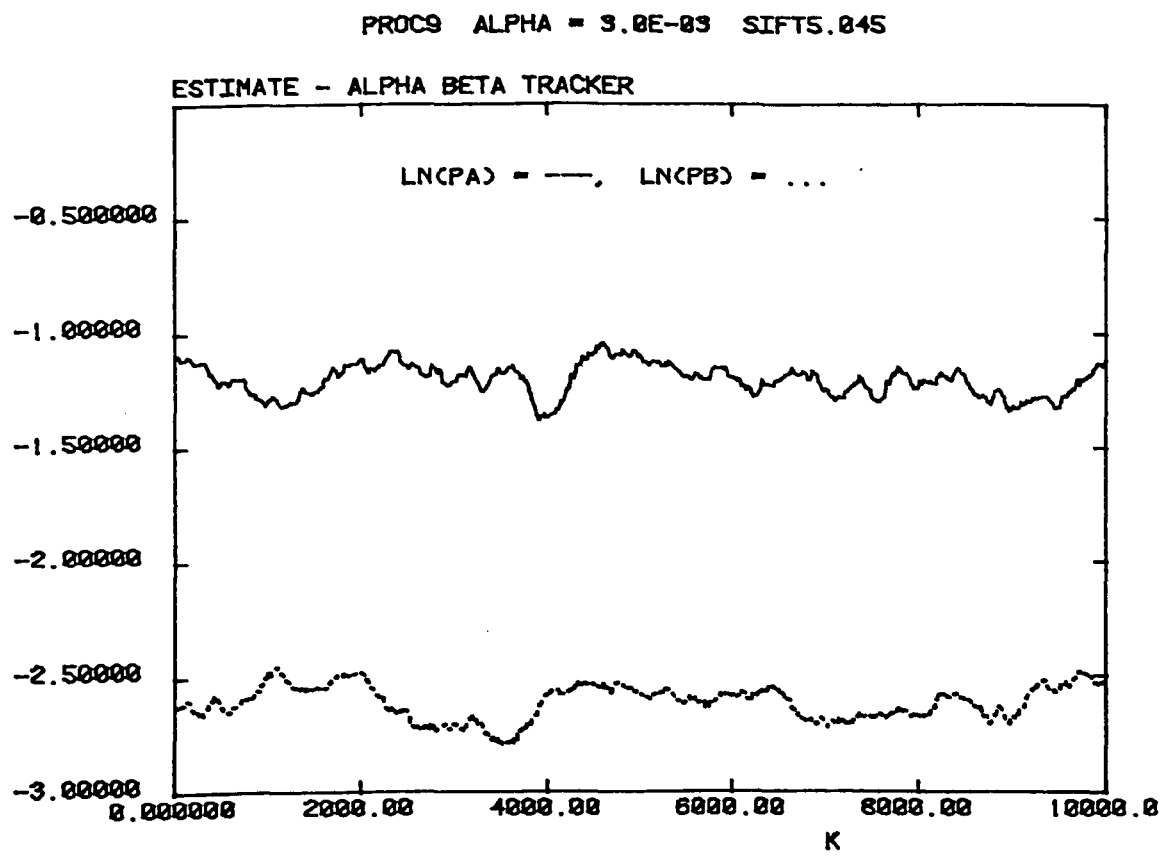


FIGURE 7.2 (e)



Estimates of the log of the Normalised Power: α β Tracker (above)

: Running Average (below)

FIGURE 7.2 (f)

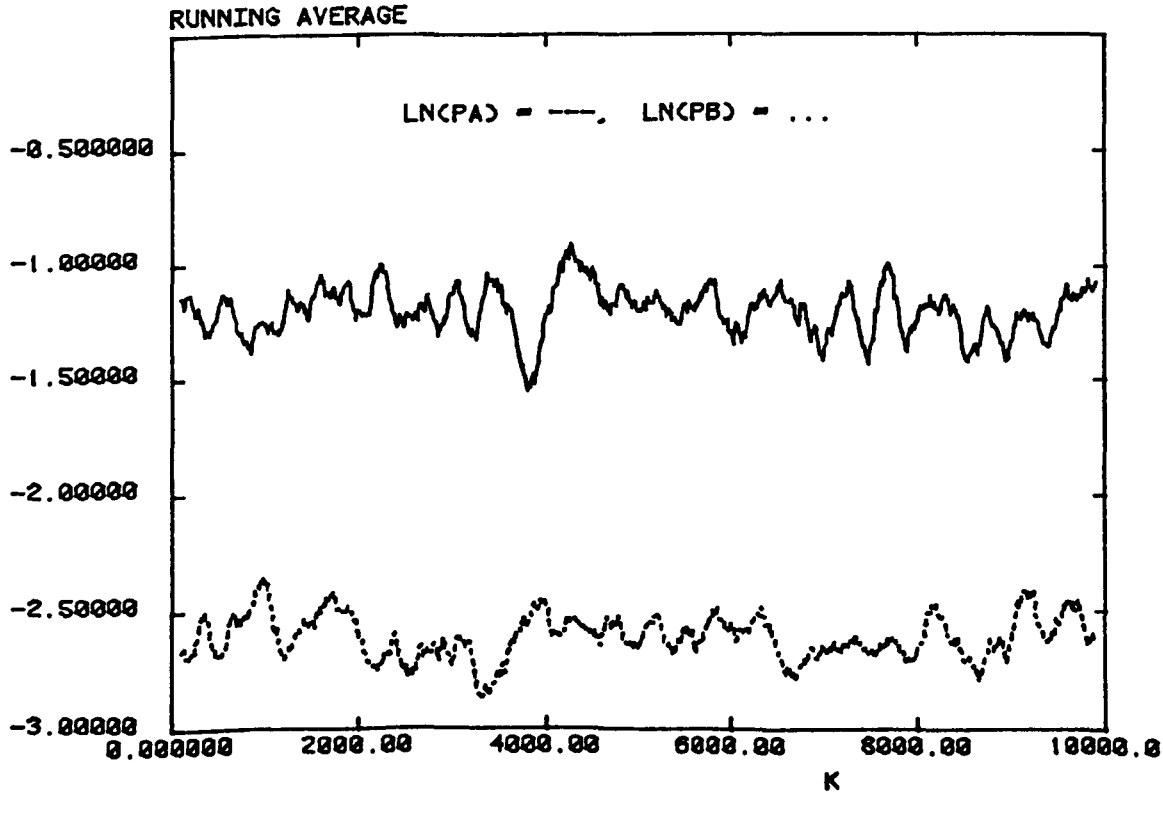
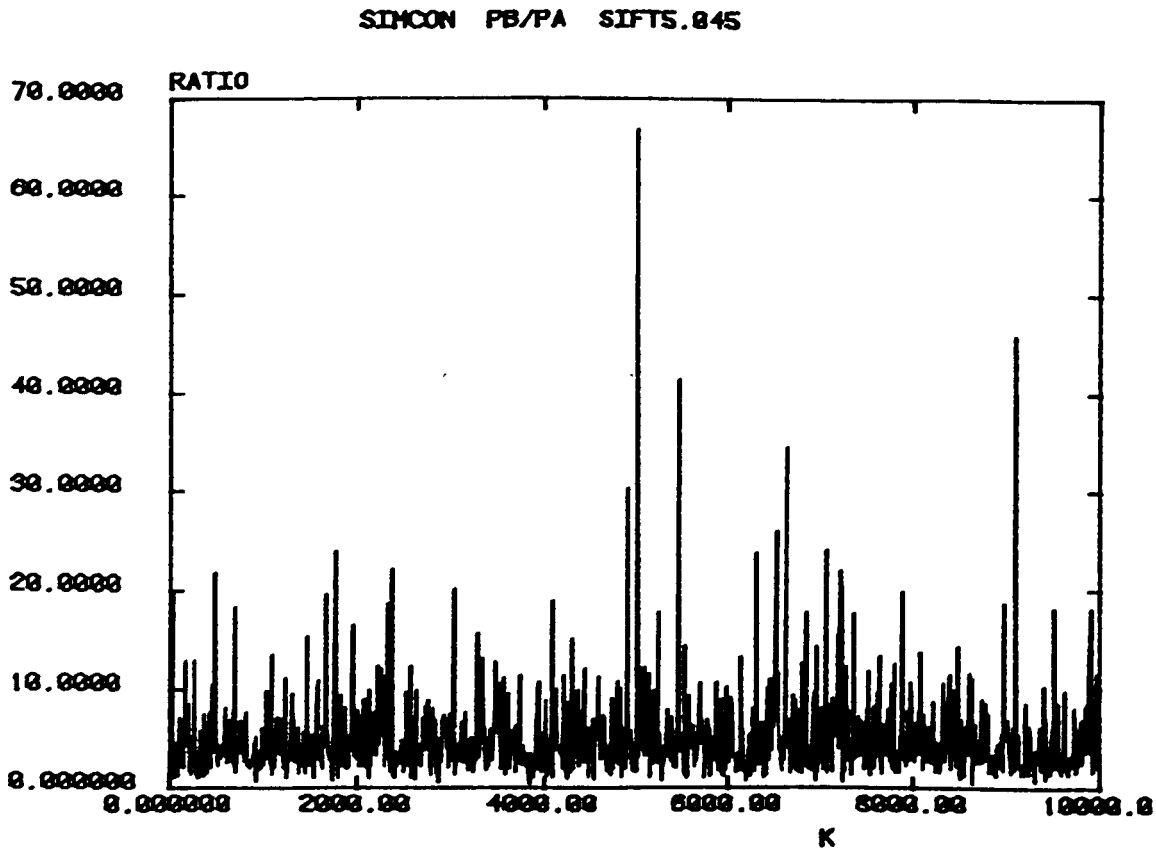


FIGURE 7.2 (g)



Ratio of the Normalised Powers in Each Channel (above)

Log-Ratio of the Normalised Powers in Each Channel (below)

FIGURE 7.2 (h)

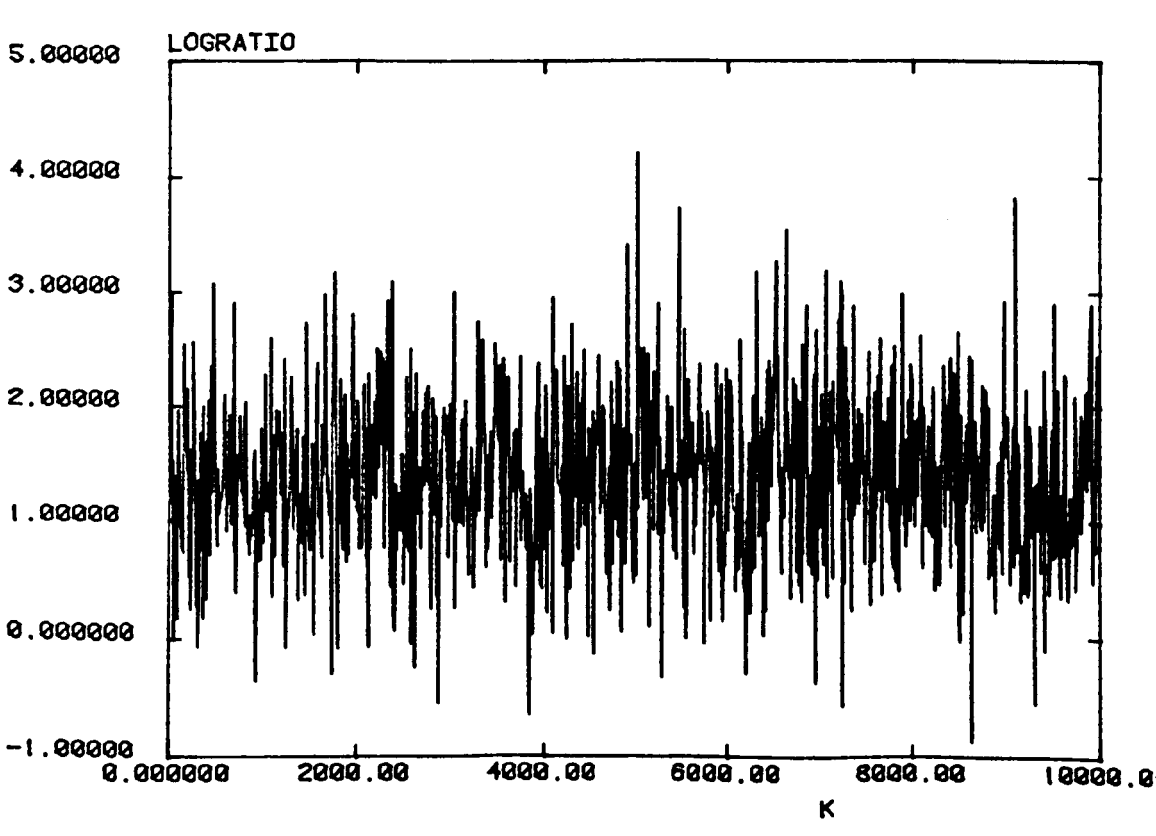
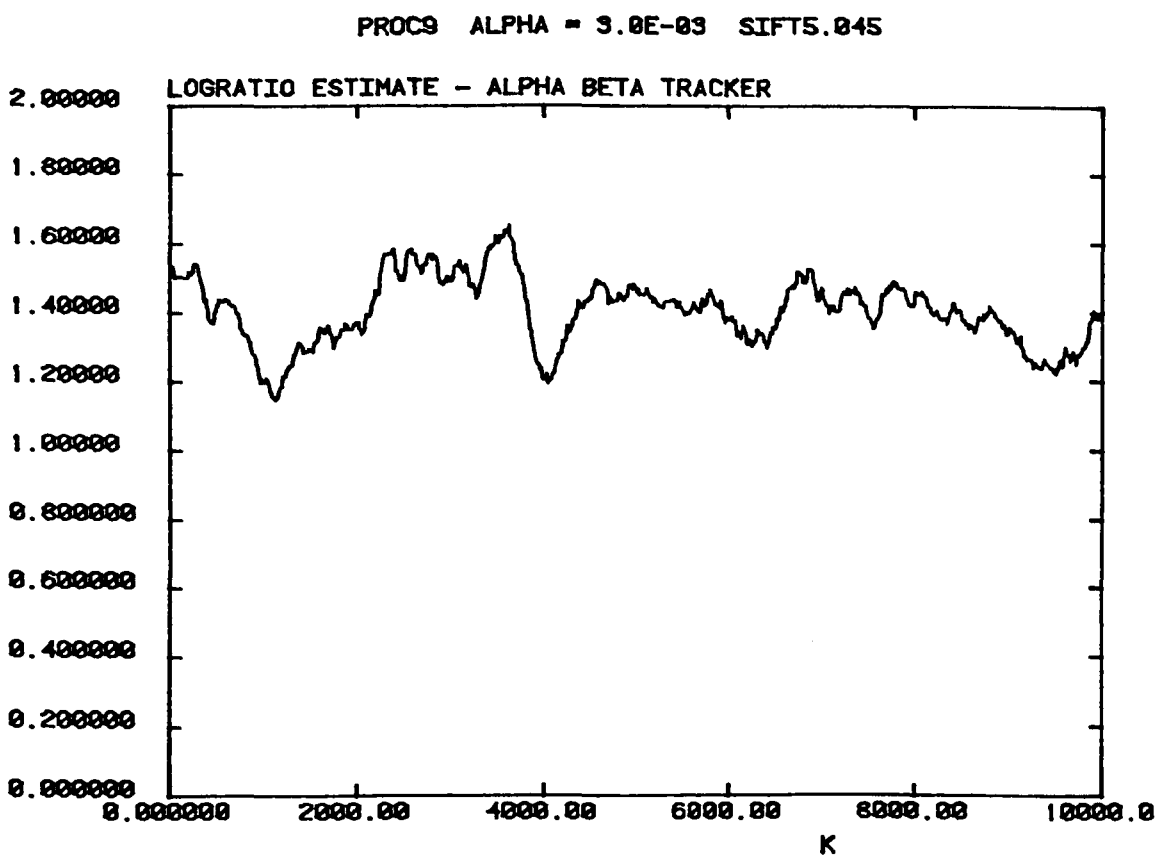


FIGURE 7.2 (i)



Log-Ratio Estimates: $\alpha\beta$ Tracker (above)

: Running Average (below)

FIGURE 7.2 (j)

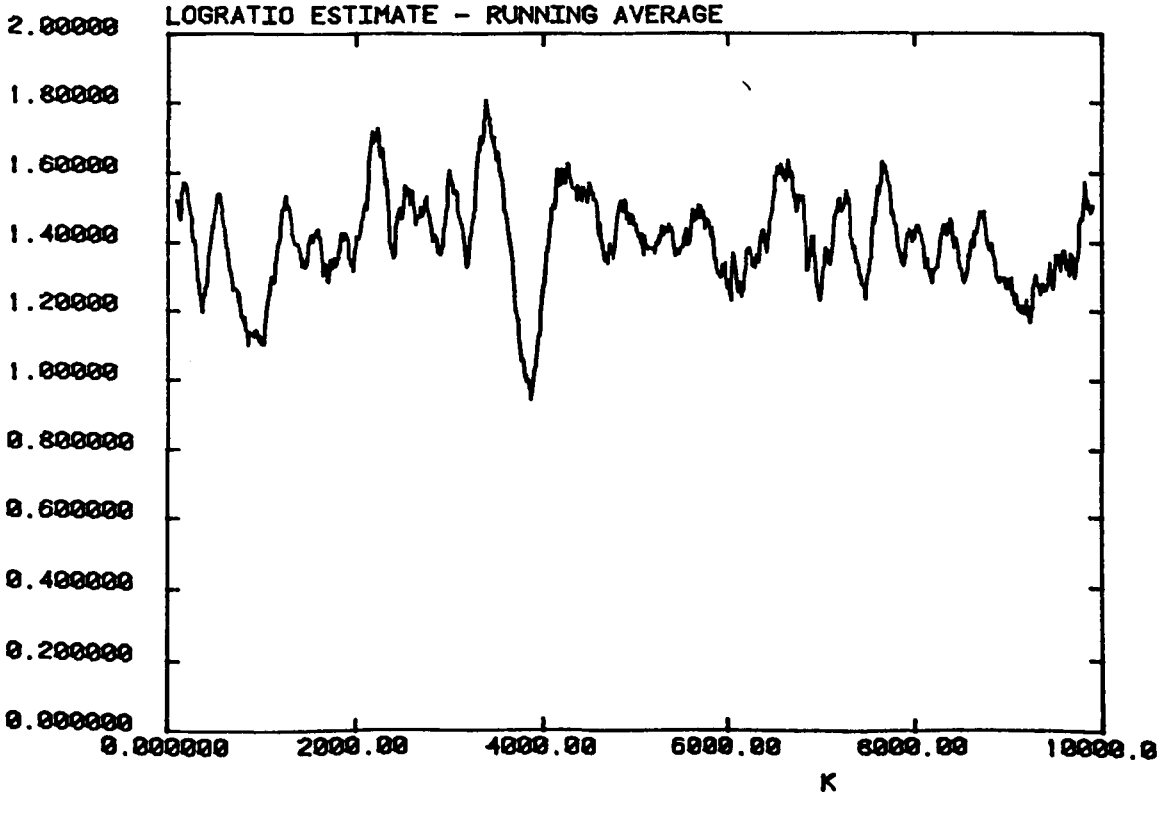
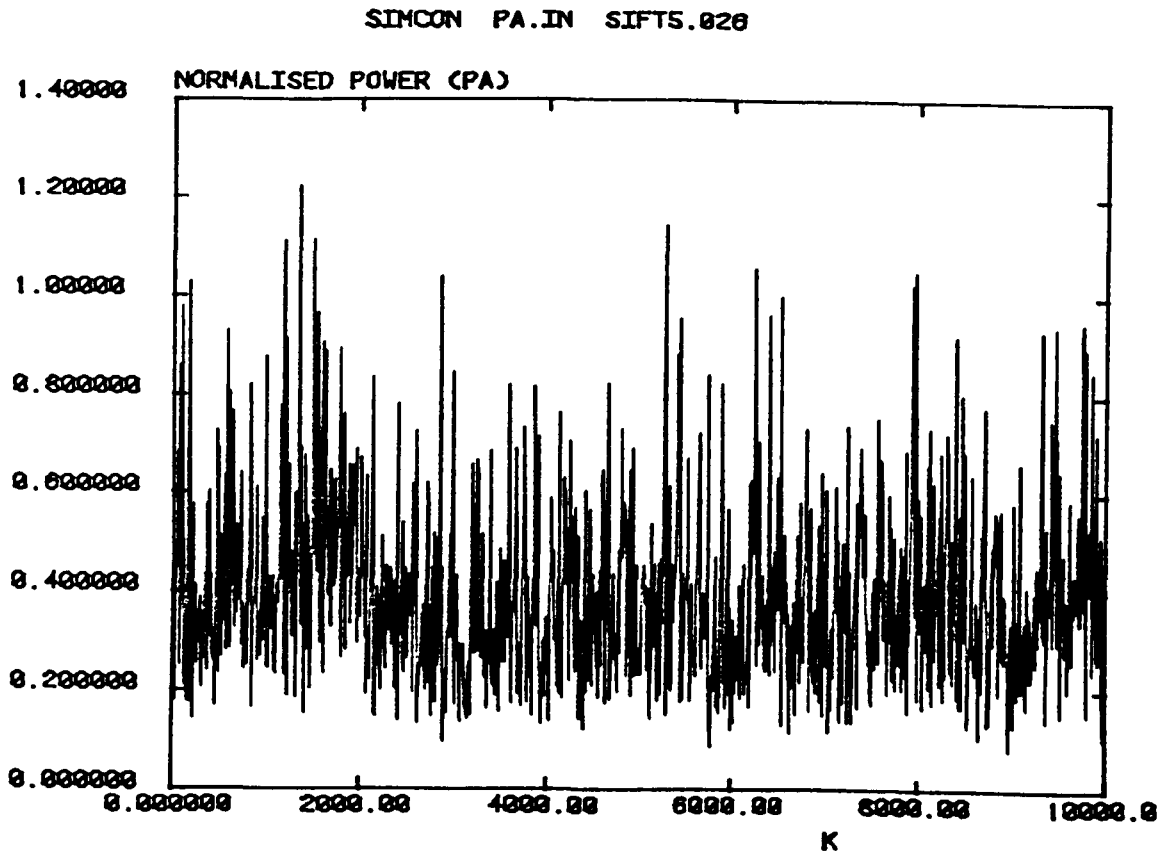


FIGURE 7.3 (a)



Normalised Power Values for the File SIFT5·026: Channel A (above)

: Channel B (below)

FIGURE 7.3 (b)

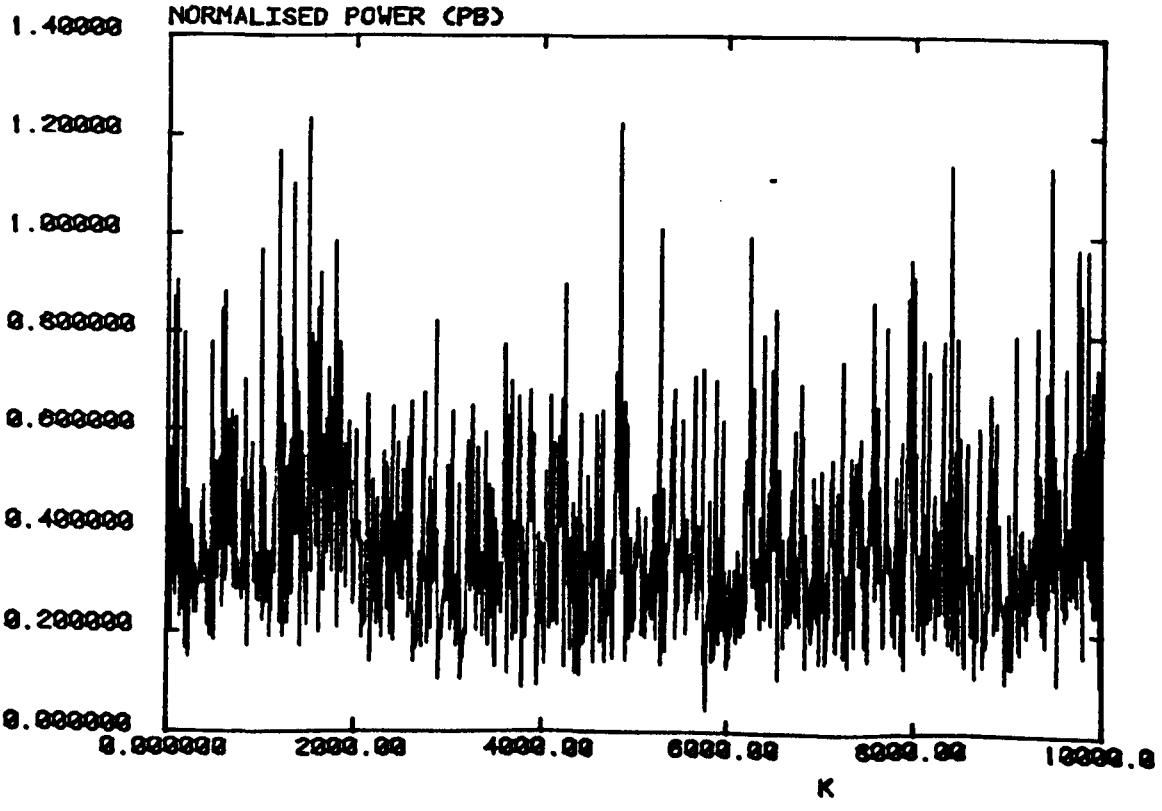
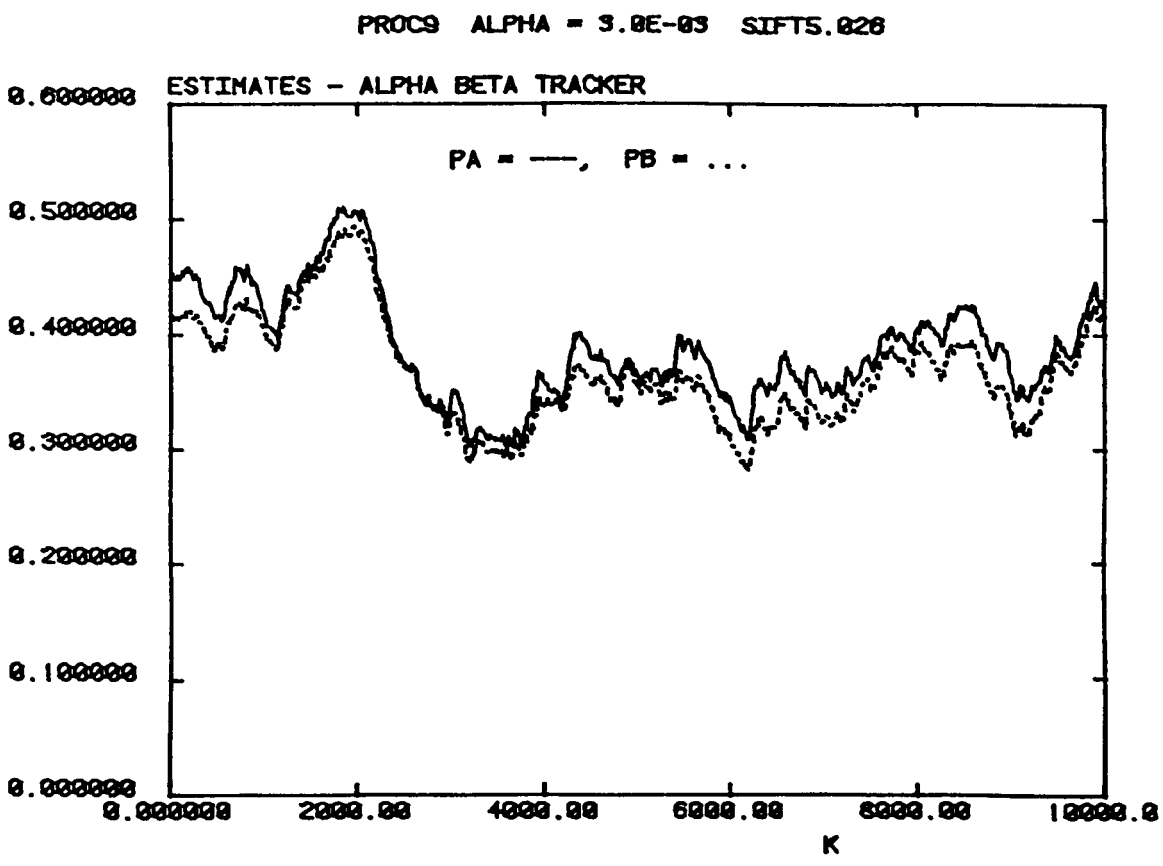


FIGURE 7.3 (c)



Normalised Power Estimates for SIFT5·026 : $\alpha\beta$ Tracker (above)

: Running Average (below)

FIGURE 7.3 (d)

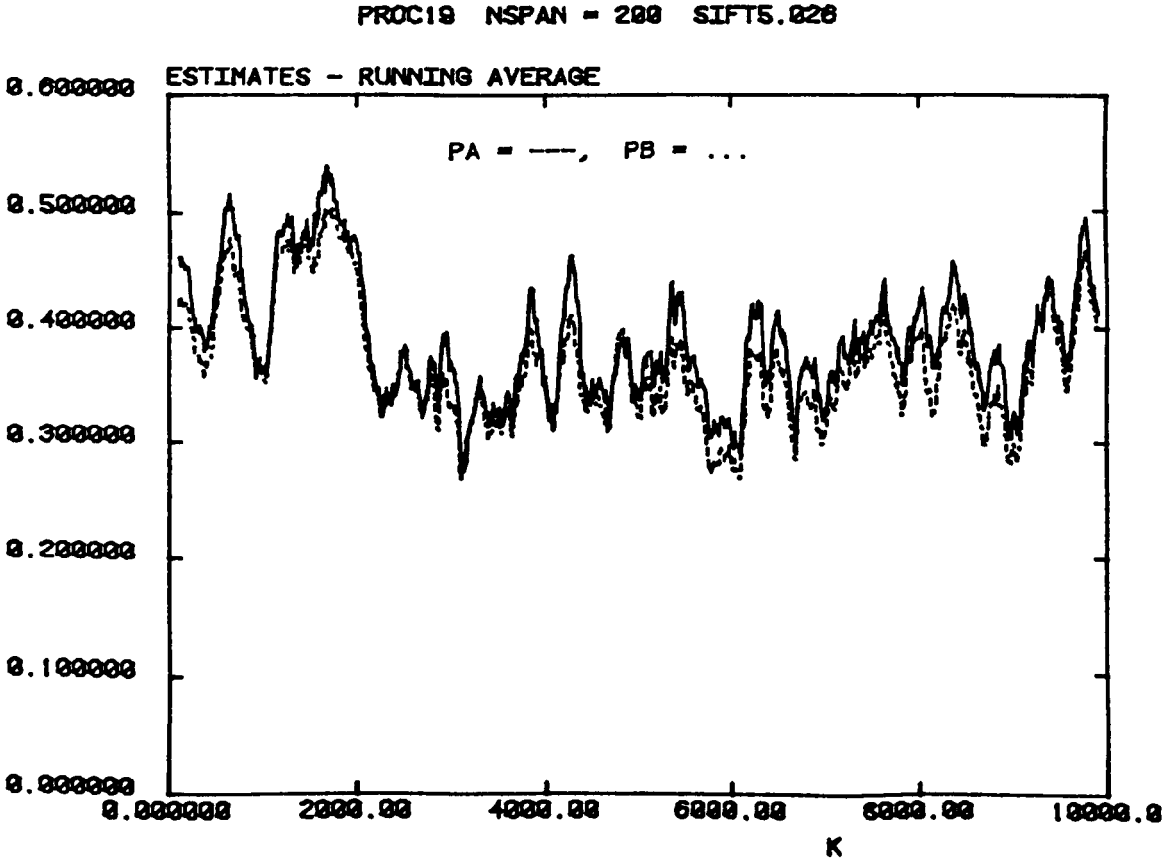
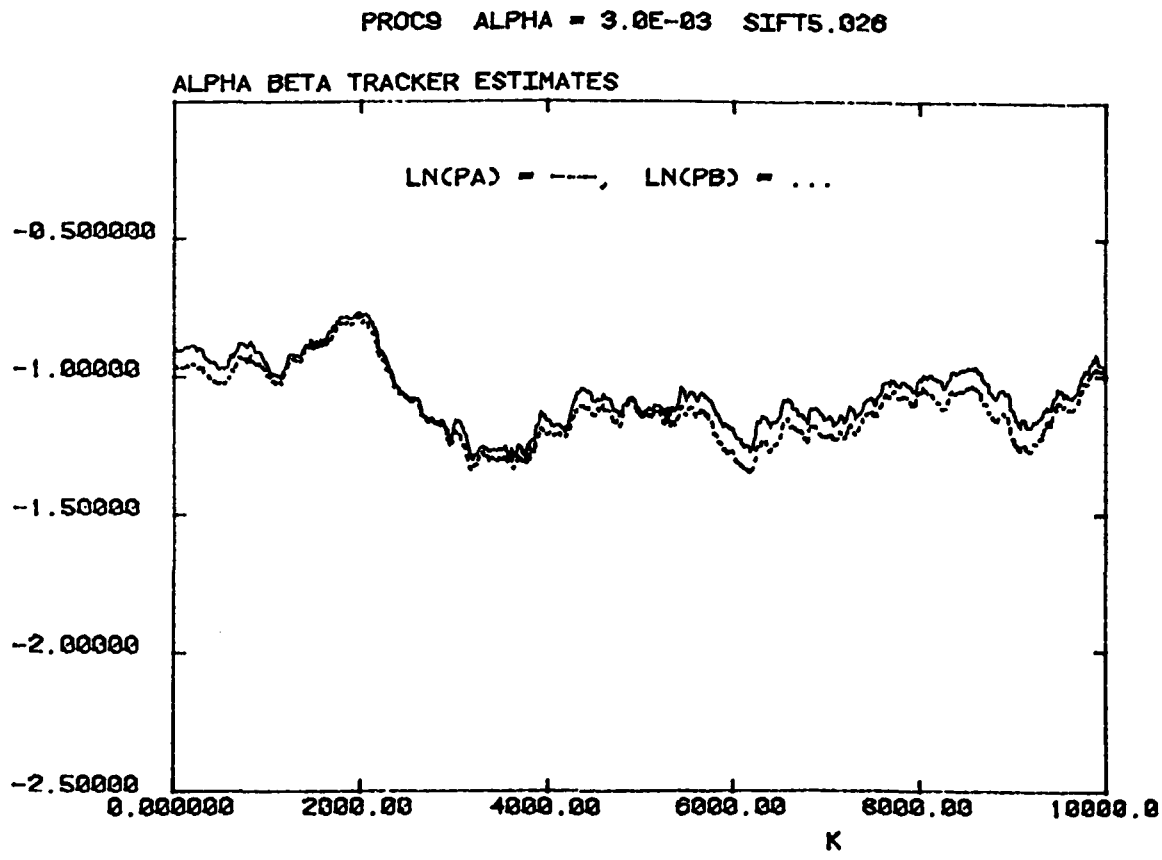


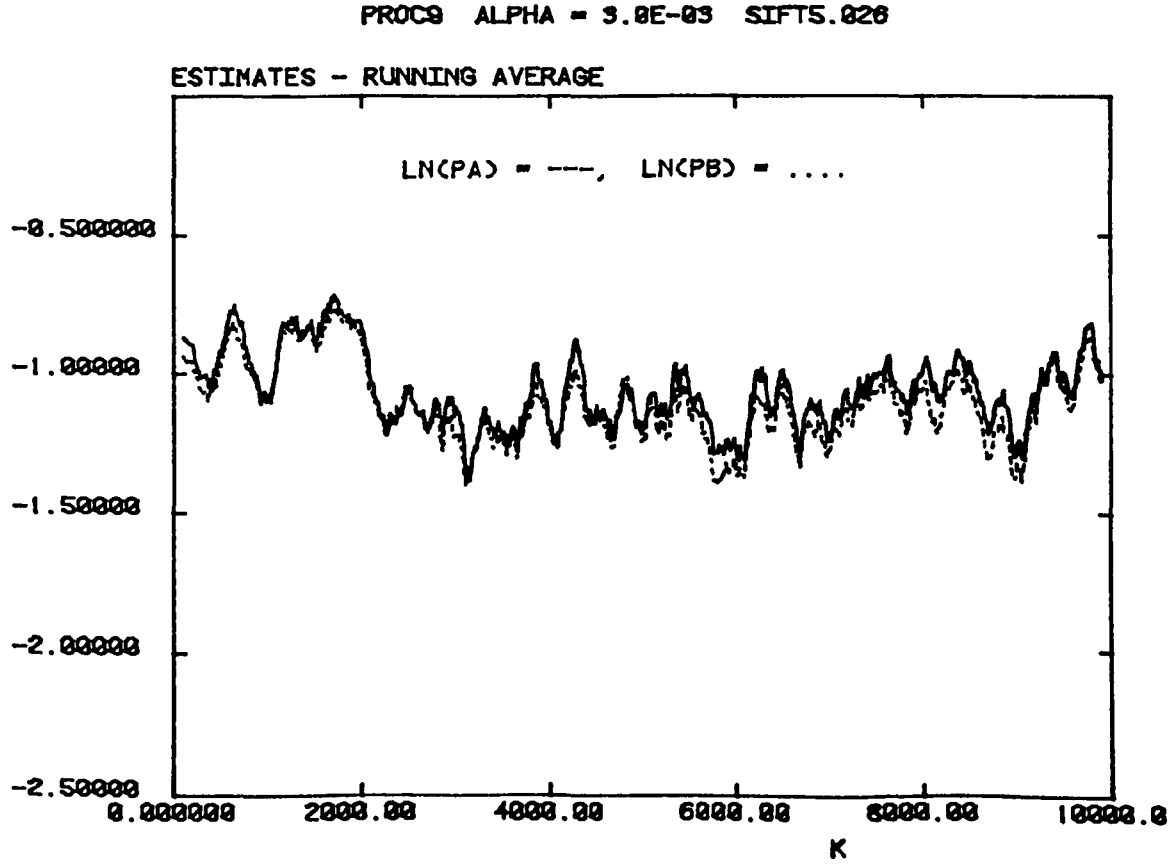
FIGURE 7.3 (e)



$\log_e$ (Normalised Power) Estimates for SIFT 5·026 : $\alpha\beta$ Tracker (above)

: Running Average (below)

FIGURE 7.3 (f)



VARIANCE REDUCTION RATIO AS FUNCTION OF ALPHA FIGURE 7.4

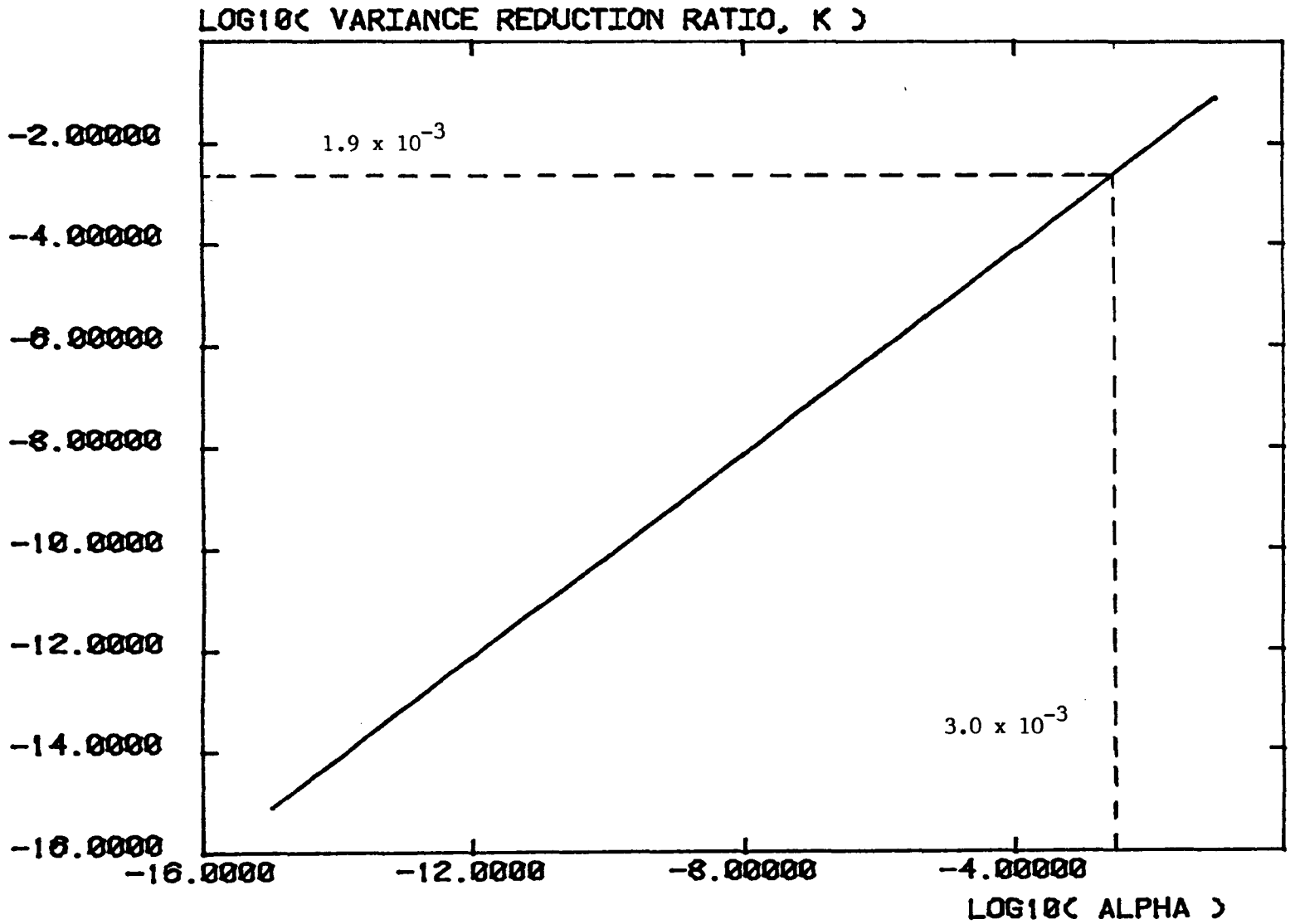
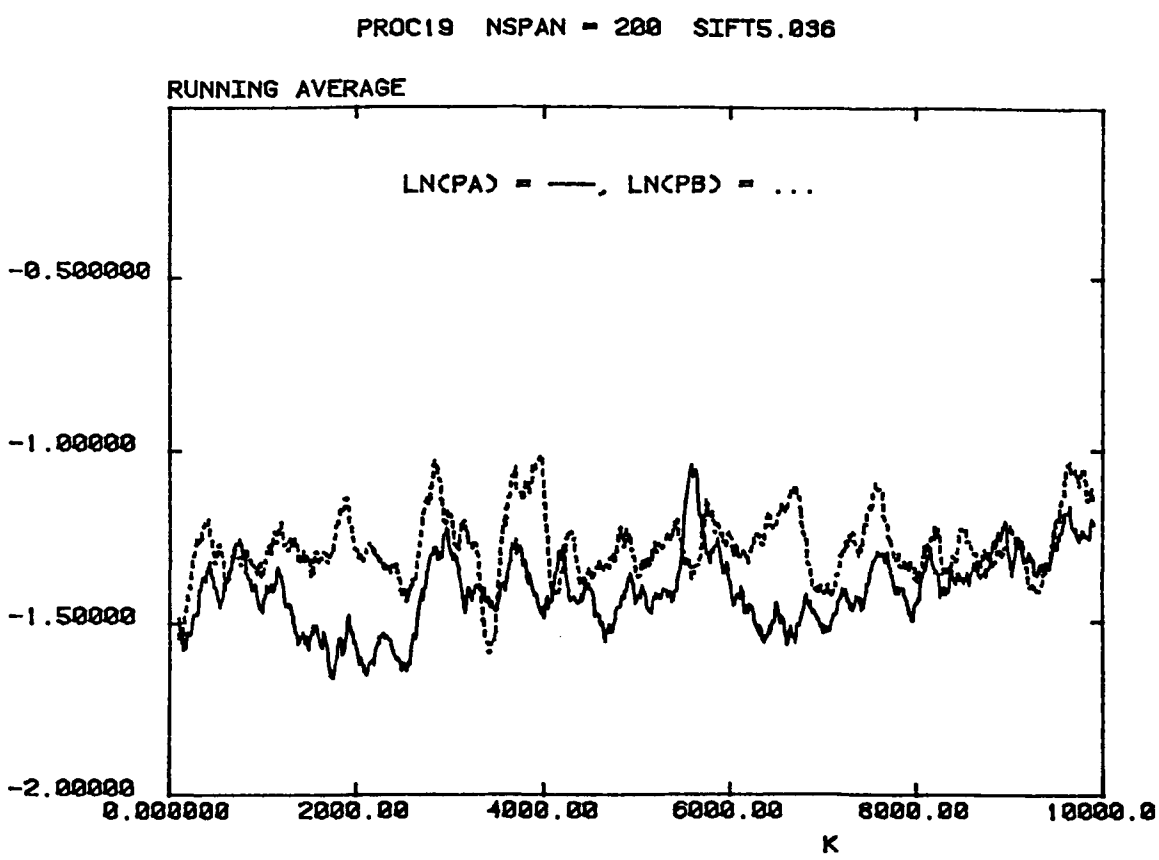


FIGURE 7.5 (a)



$\log_e$ (Normalised Power) estimates: for SIFT 5·036 (above)

: for SIFT 5·037 (below)

FIGURE 7.5 (b)

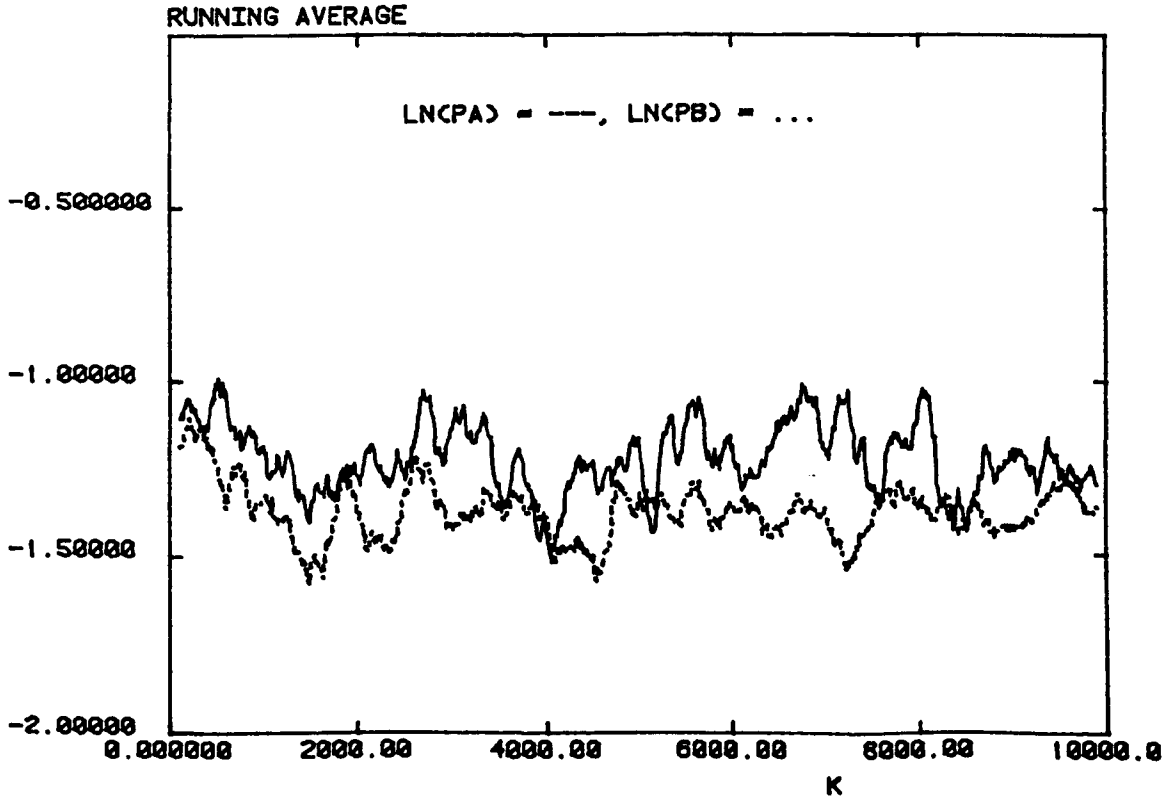
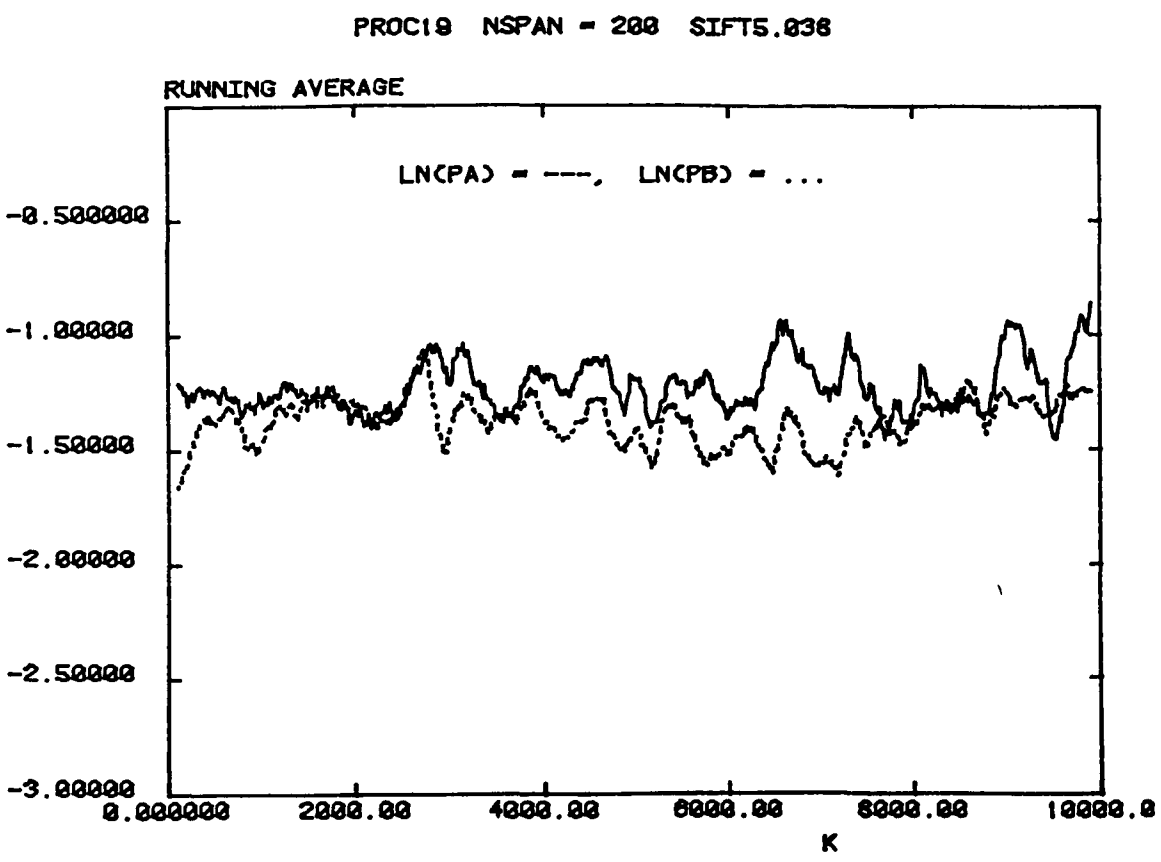


FIGURE 7.5 (c)



$\log_e$ (Normalised Power) Estimates: for SIFT5·038 (above)

: for SIFT5·043 (below)

FIGURE 7.5 (d)

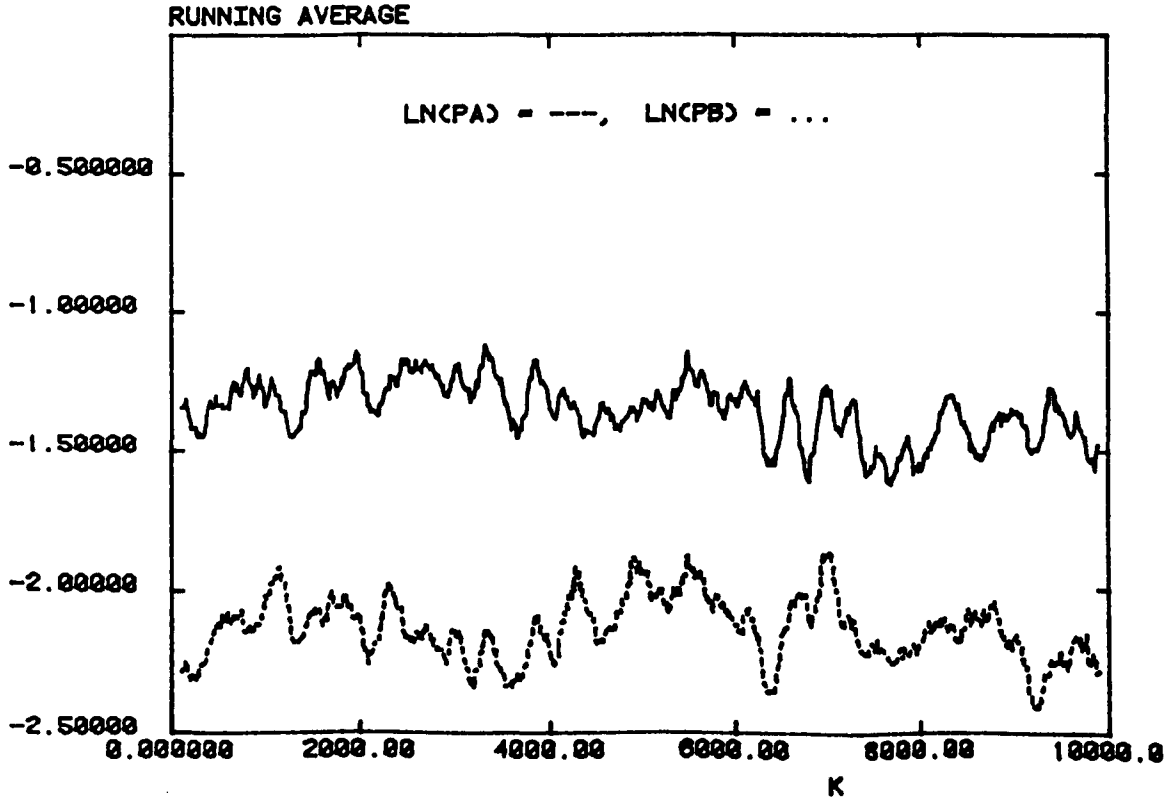
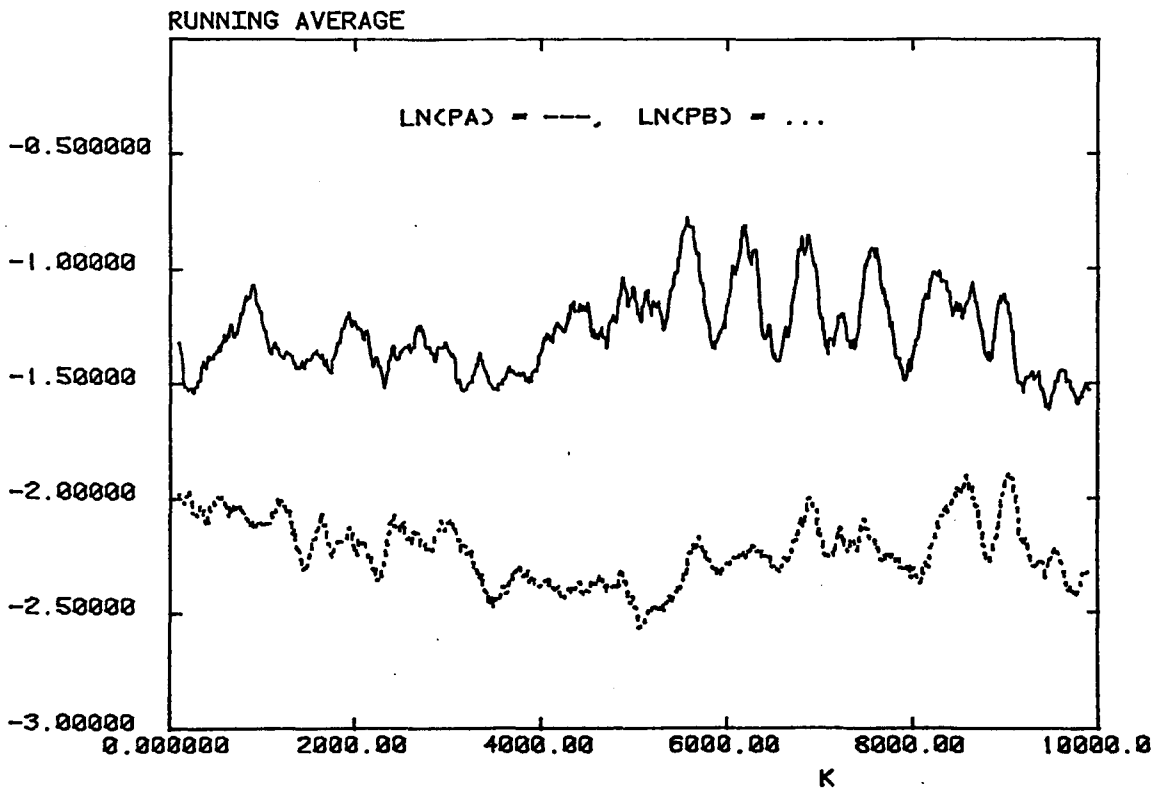


FIGURE 7.5 (e)



$\log_e$ (Normalised Power) Estimates for SIFT5·044 : NSPAN = 200 (above)

: NSPAN = 50 (below)

FIGURE 7.5 (f)

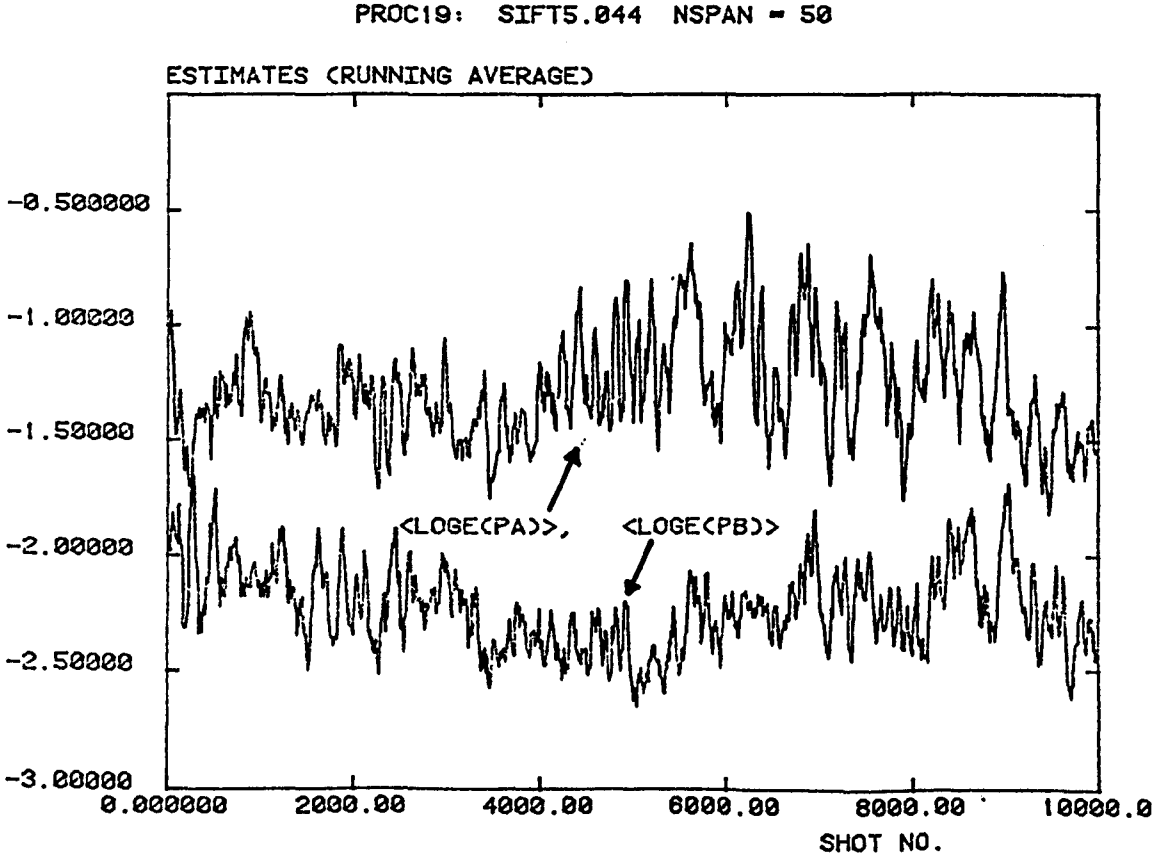
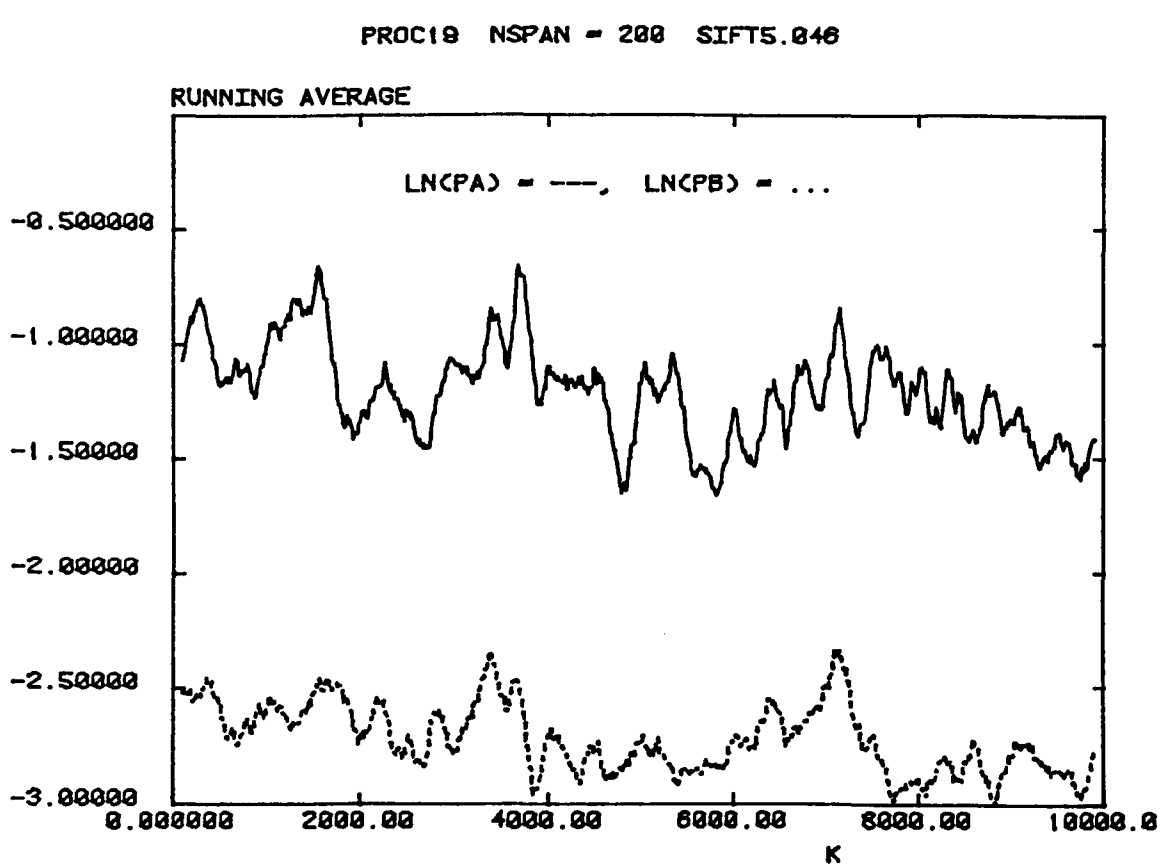


FIGURE 7.5 (g)



Log_e (Normalised Power) Estimates: for SIFT 5·046 (above)

: for SIFT 5·050 (below)

FIGURE 7.5 (h)

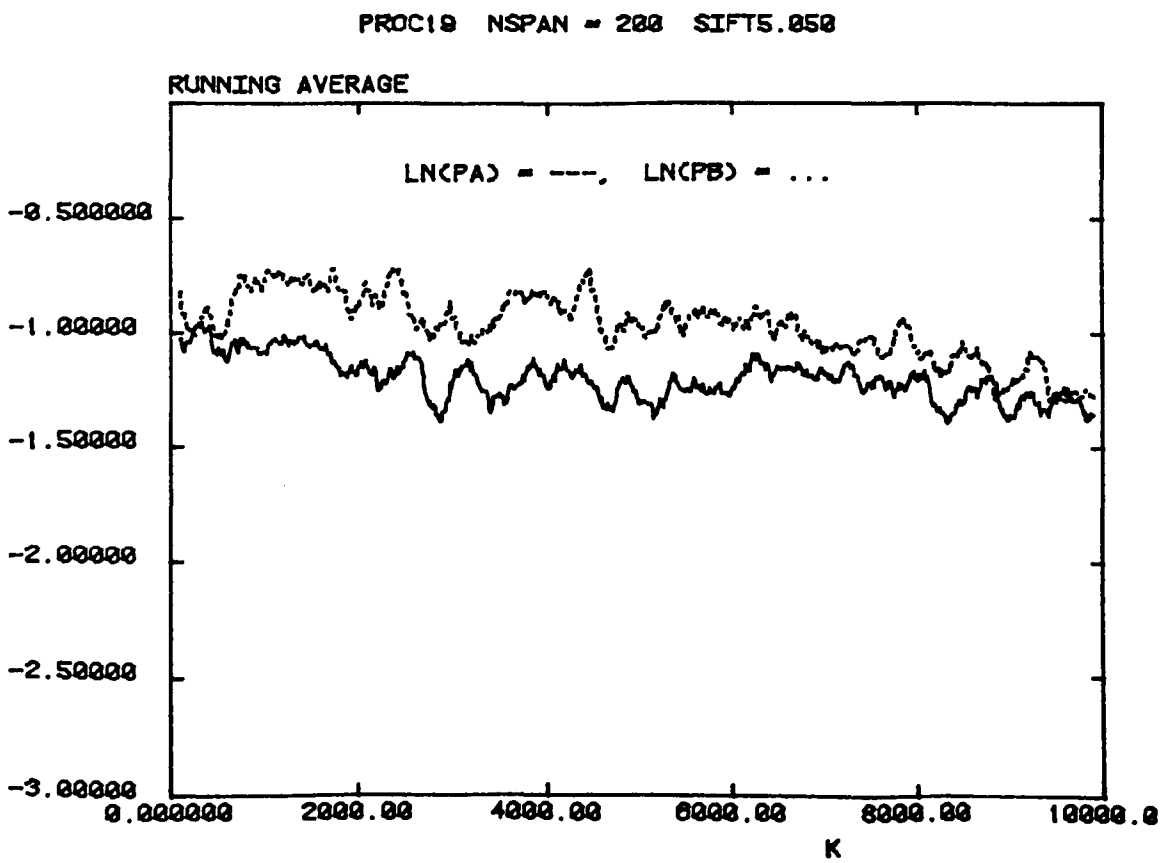
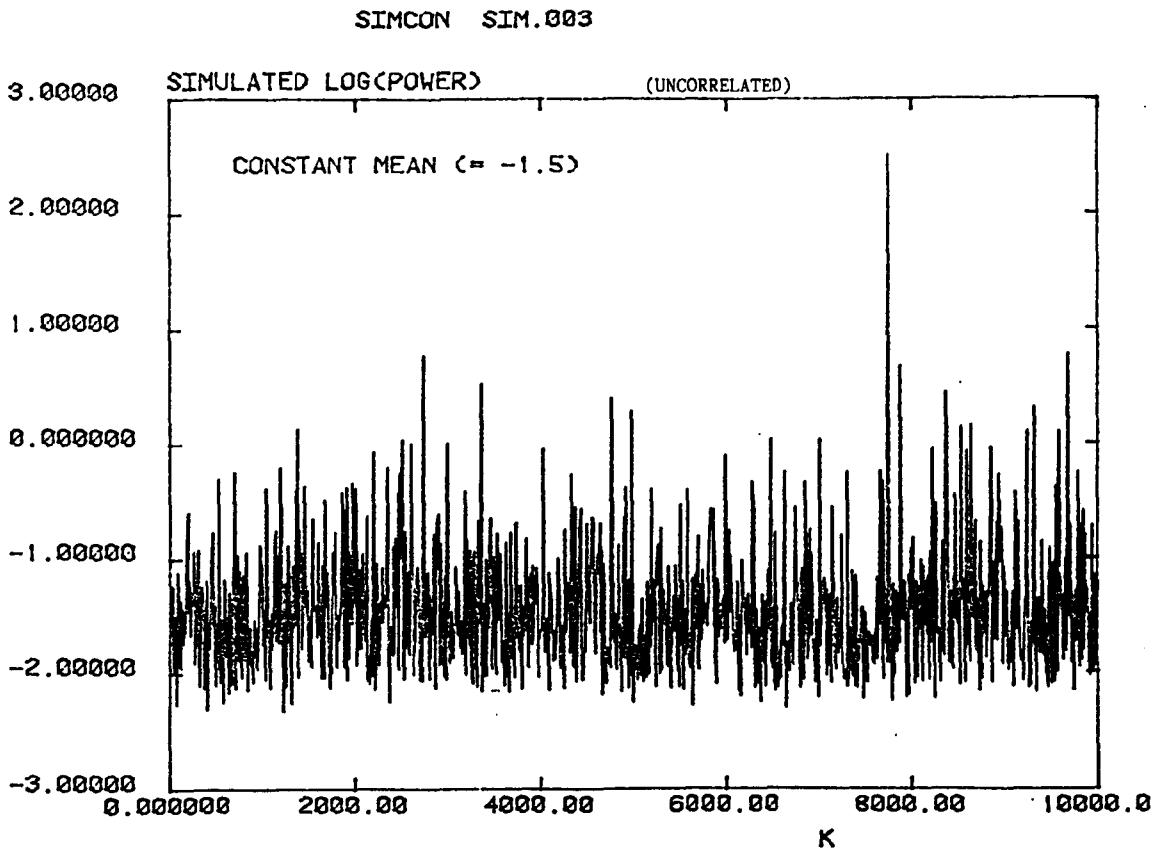


FIGURE 7.6 (a)



Simulated $\log_e$ (Normalised Power) Data : Uncorrelated Speckle (above)

: Correlated Speckle (below)

FIGURE 7.6 (b)

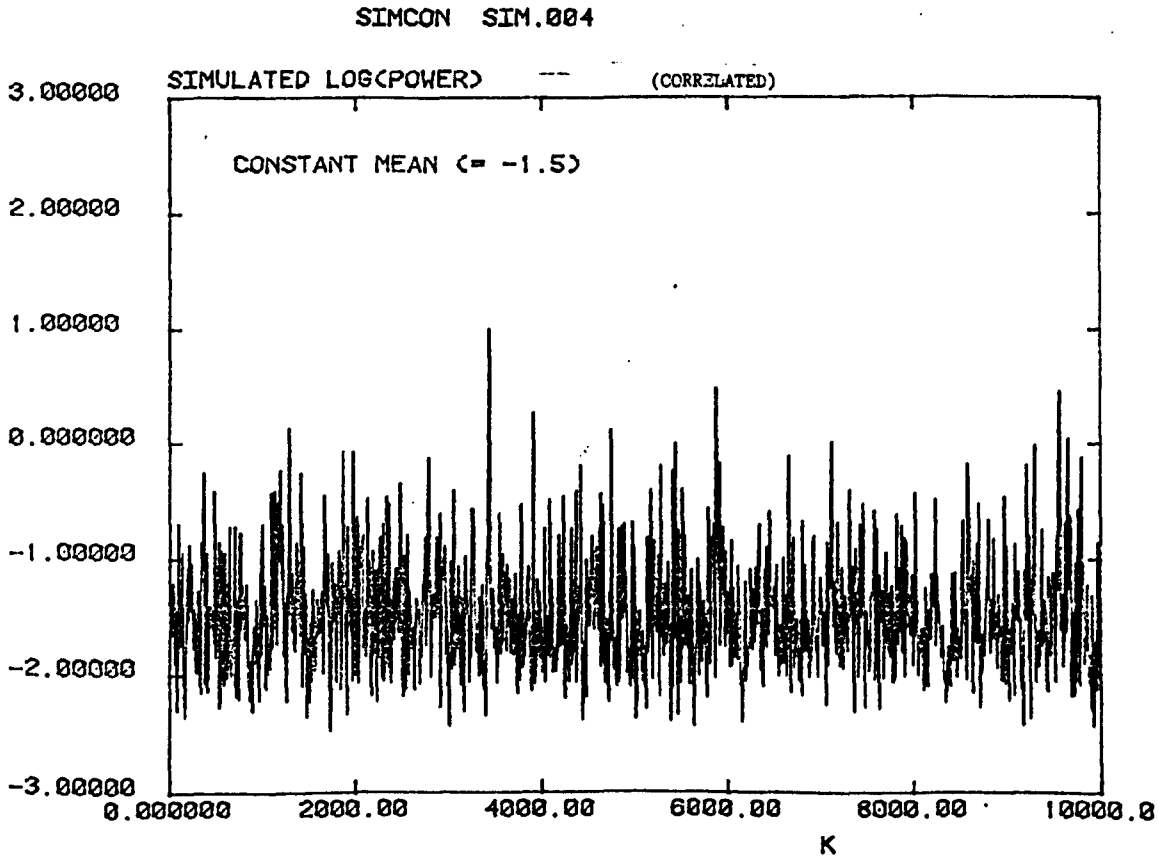
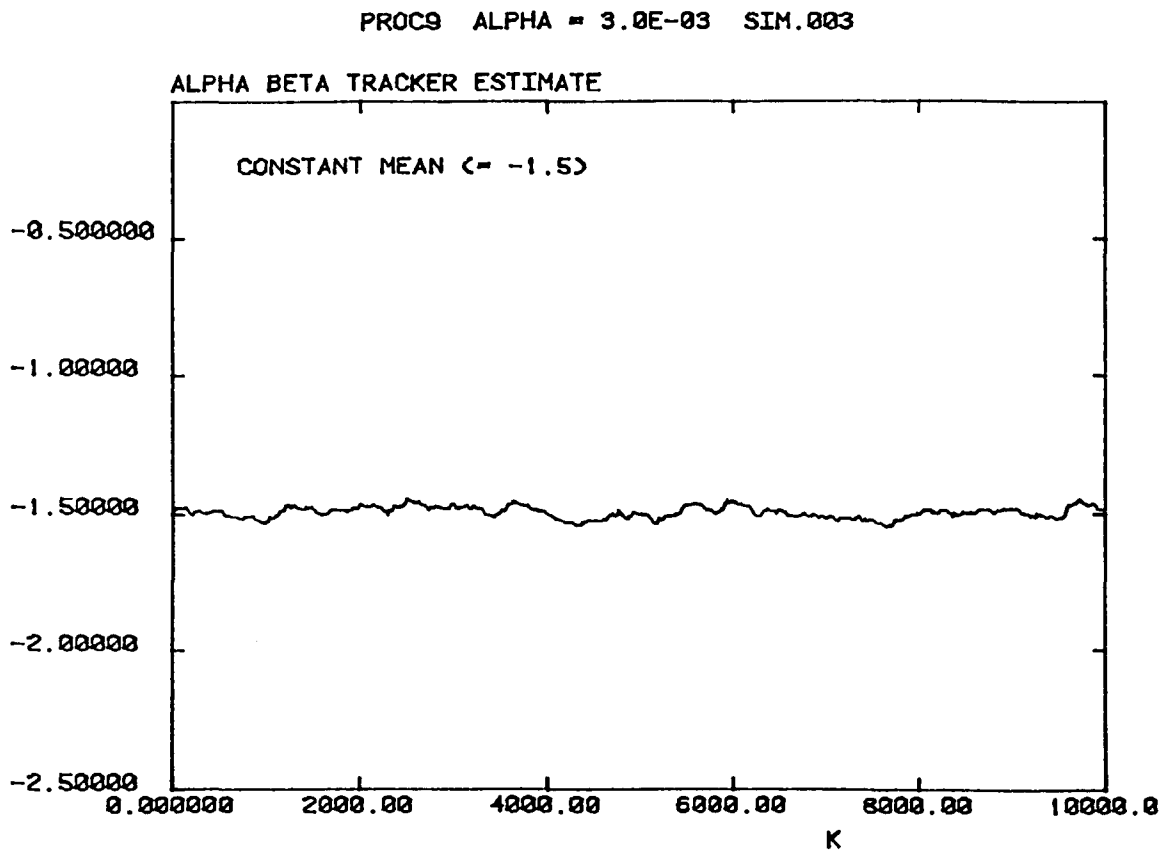


FIGURE 7.6 (c)



α β Tracker Estimates of Simulated Data: Uncorrelated Speckle (above)

: Correlated Speckle (below)

FIGURE 7.6 (d)

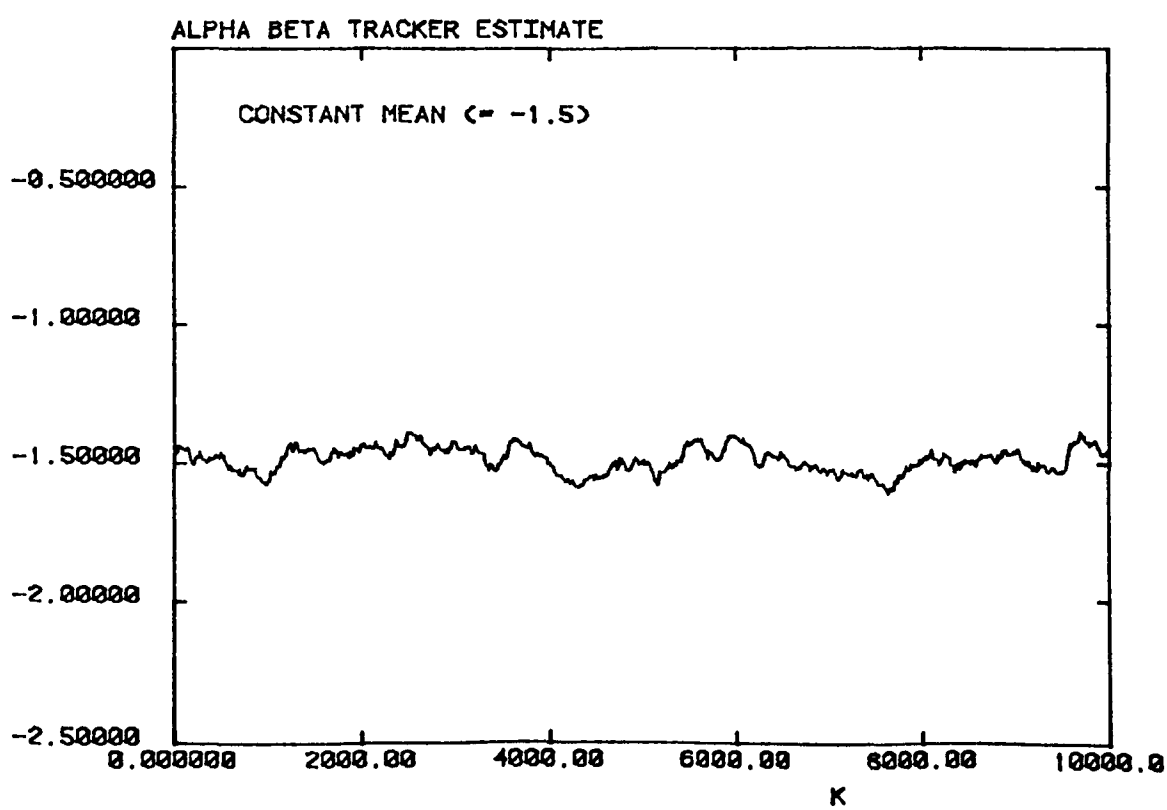
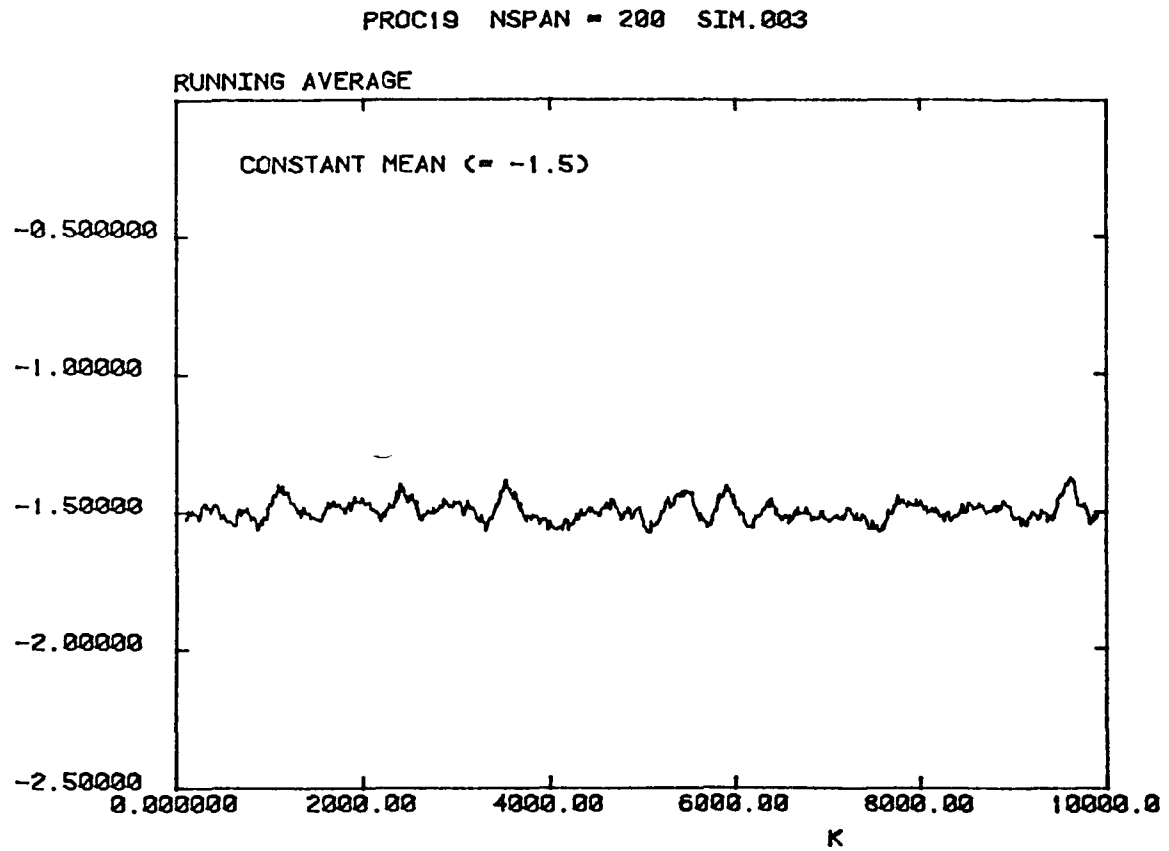


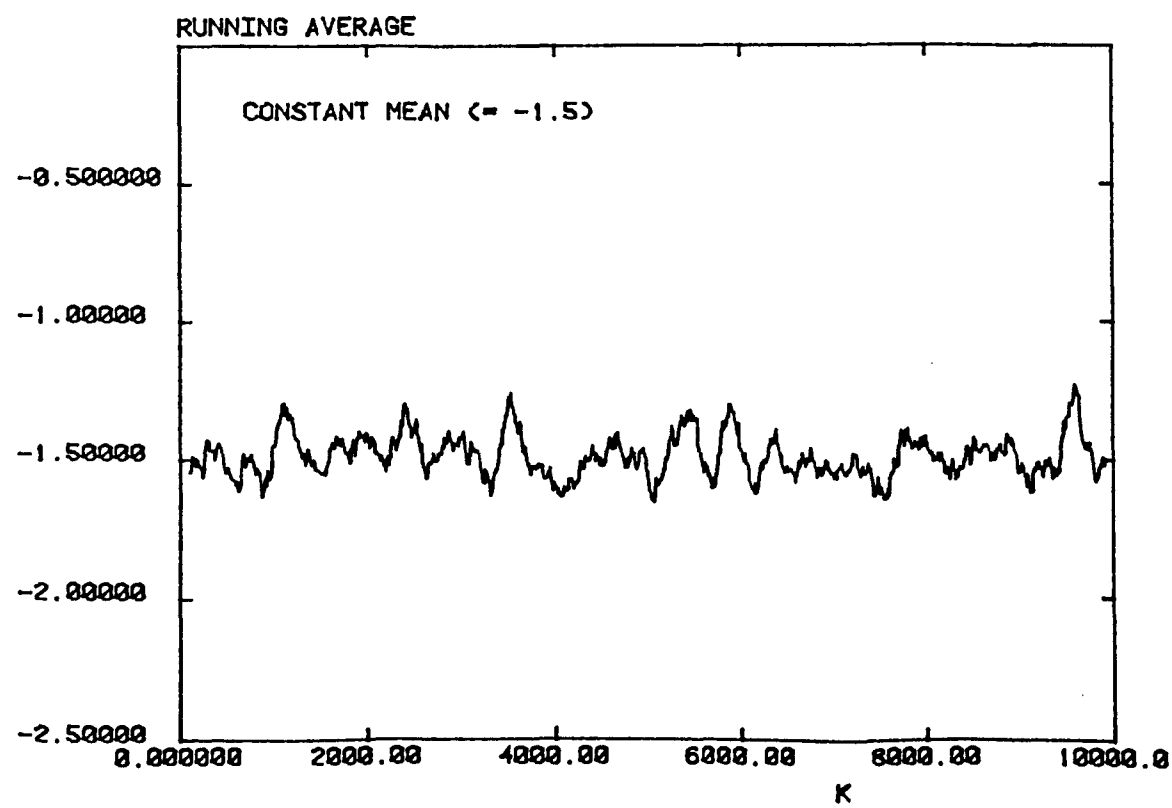
FIGURE 7.6 (e)



Running Average Estimate of Simulated Data: Uncorrelated Speckle (above)

: Correlated Speckle (below)

FIGURE 7.6 (f)



Simulated Signal

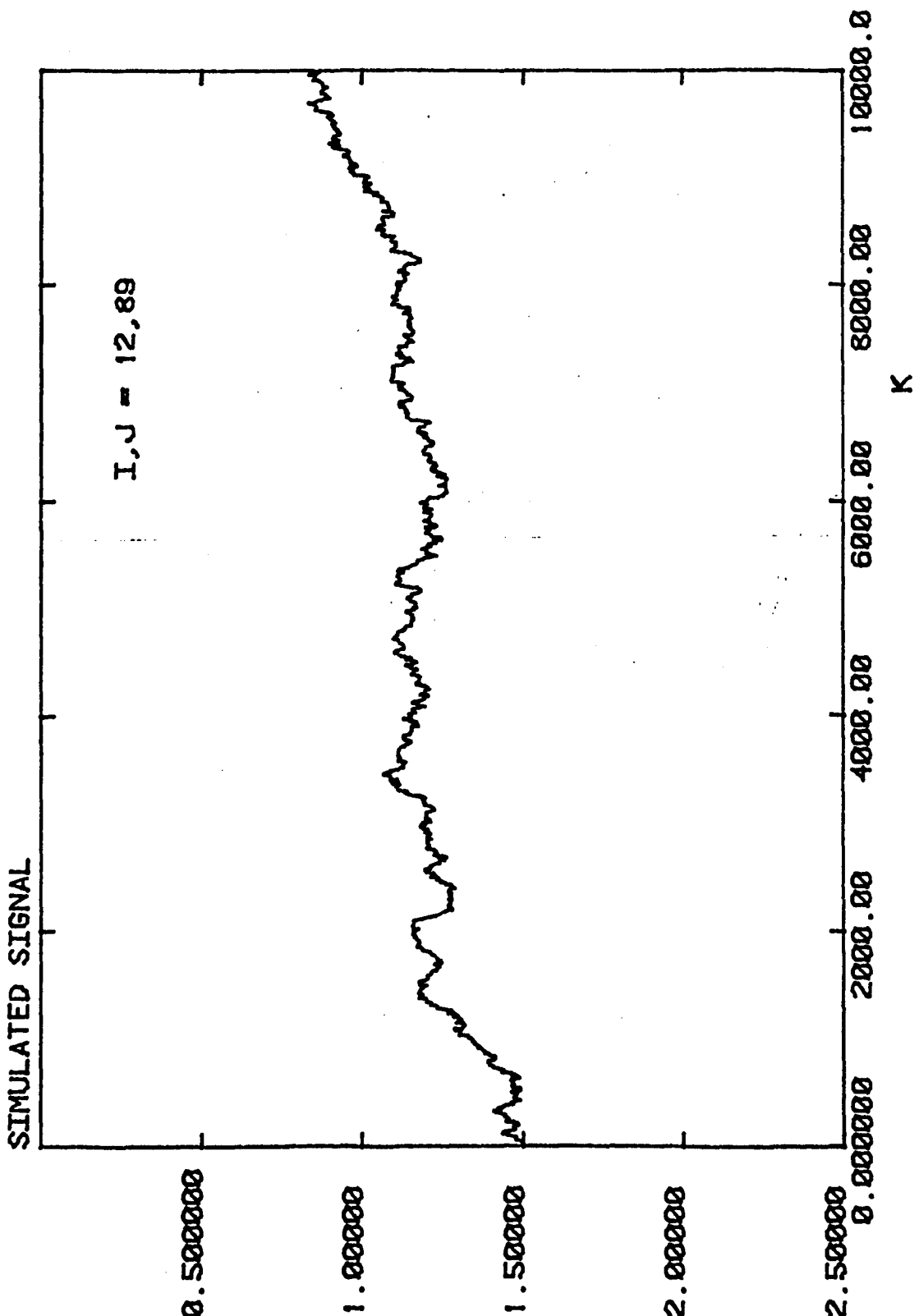
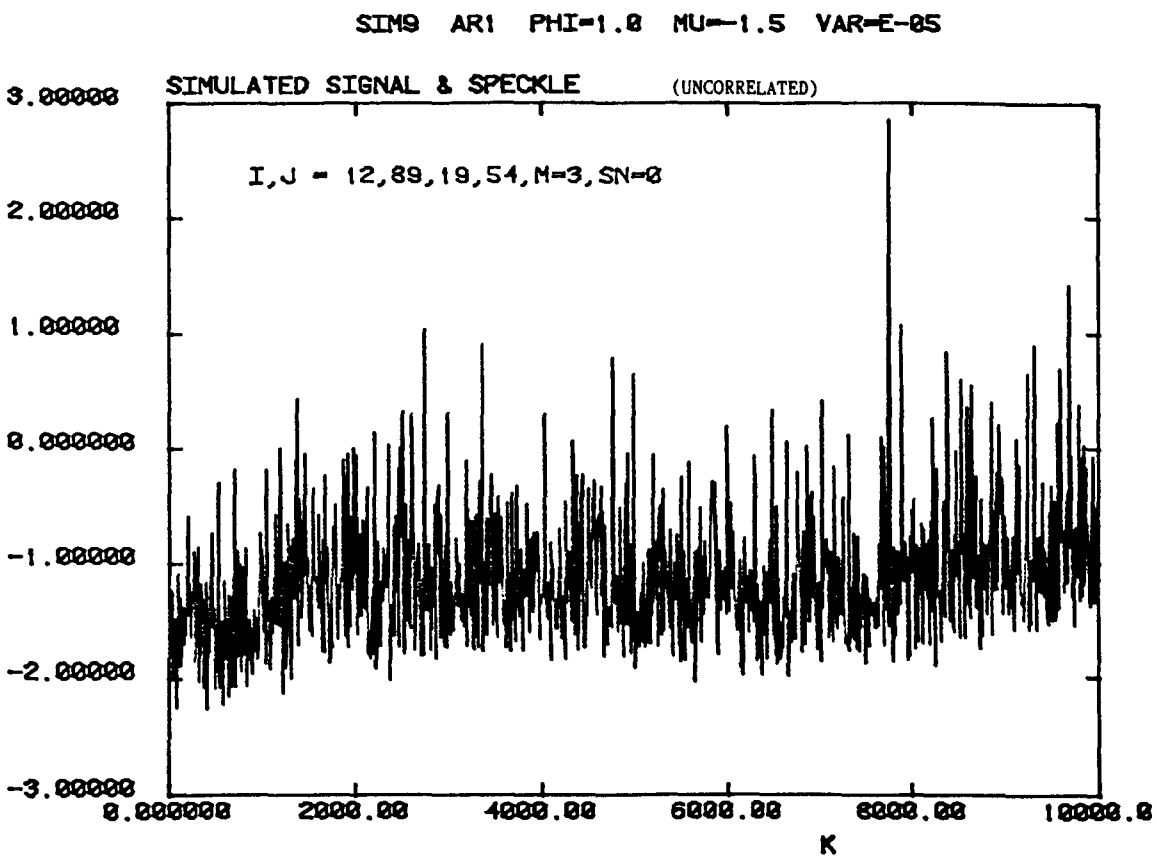


FIGURE 7.7 (a)

FIGURE 7.7 (b)



Simulated Signal : with Uncorrelated Speckle (above)

: with Correlated Speckle (below)

FIGURE 7.7 (c)

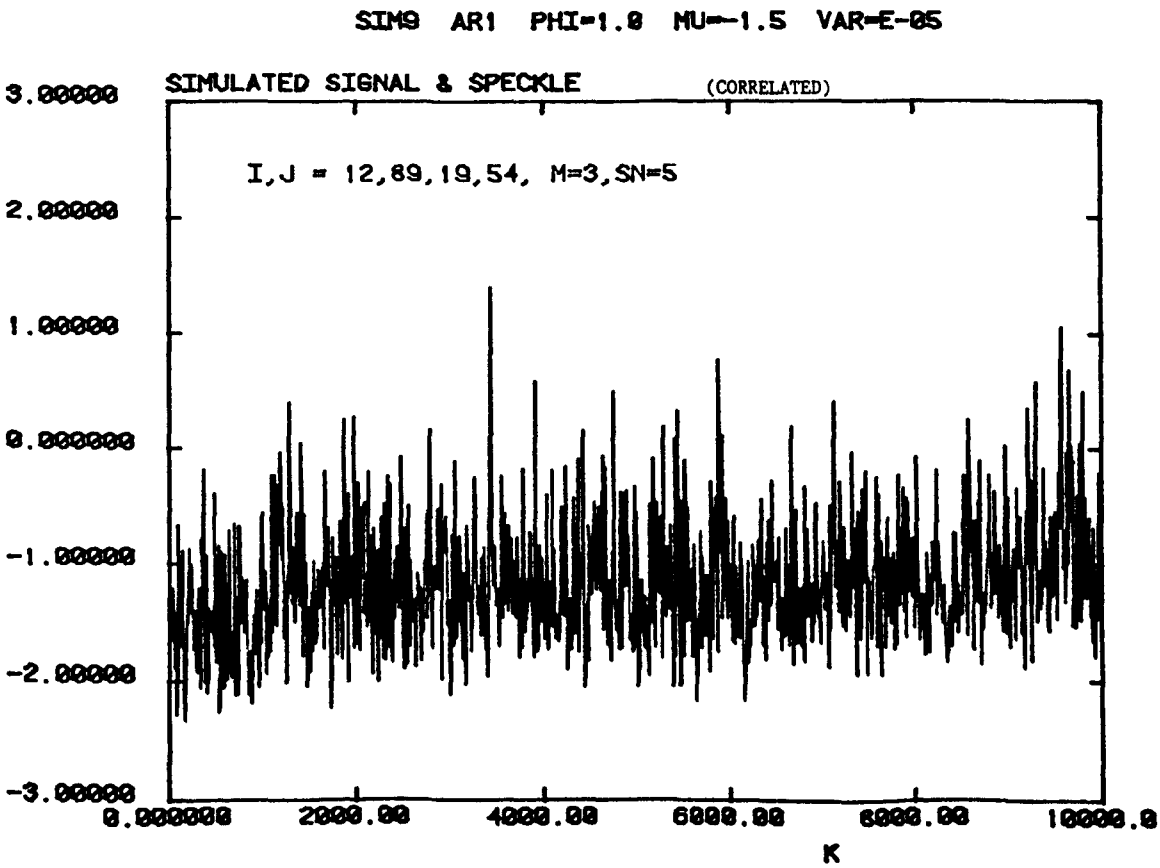
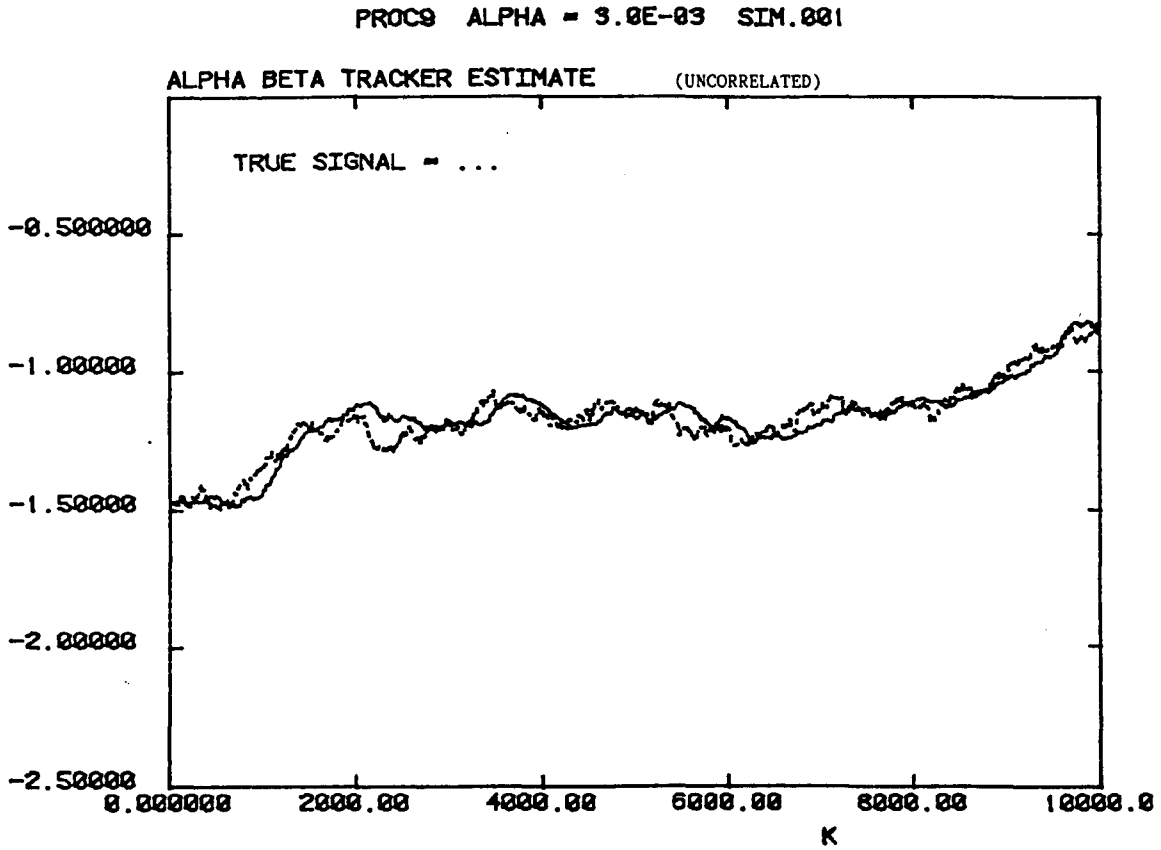


FIGURE 7.7 (d)



$\alpha\beta$ Tracker Estimates of Simulated Data: Uncorrelated Speckle (above)

: Correlated Speckle (below)

FIGURE 7.7 (e)

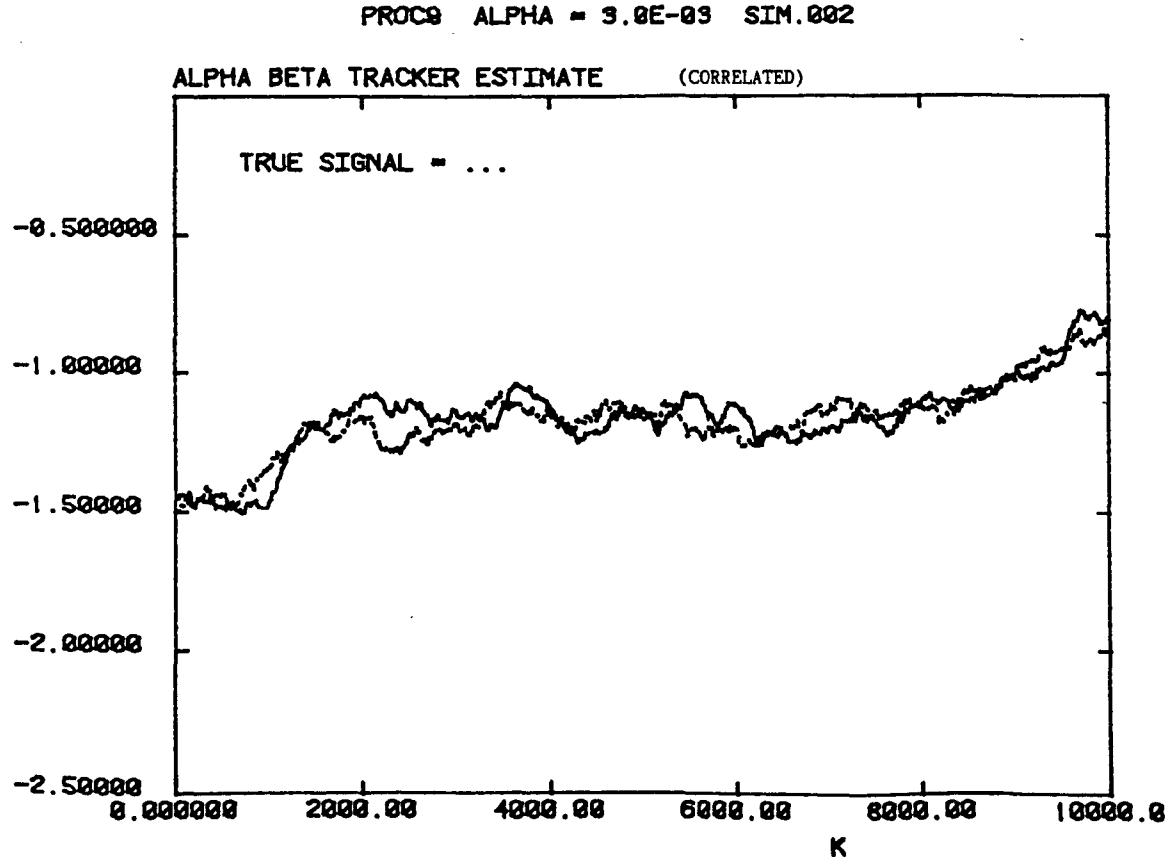
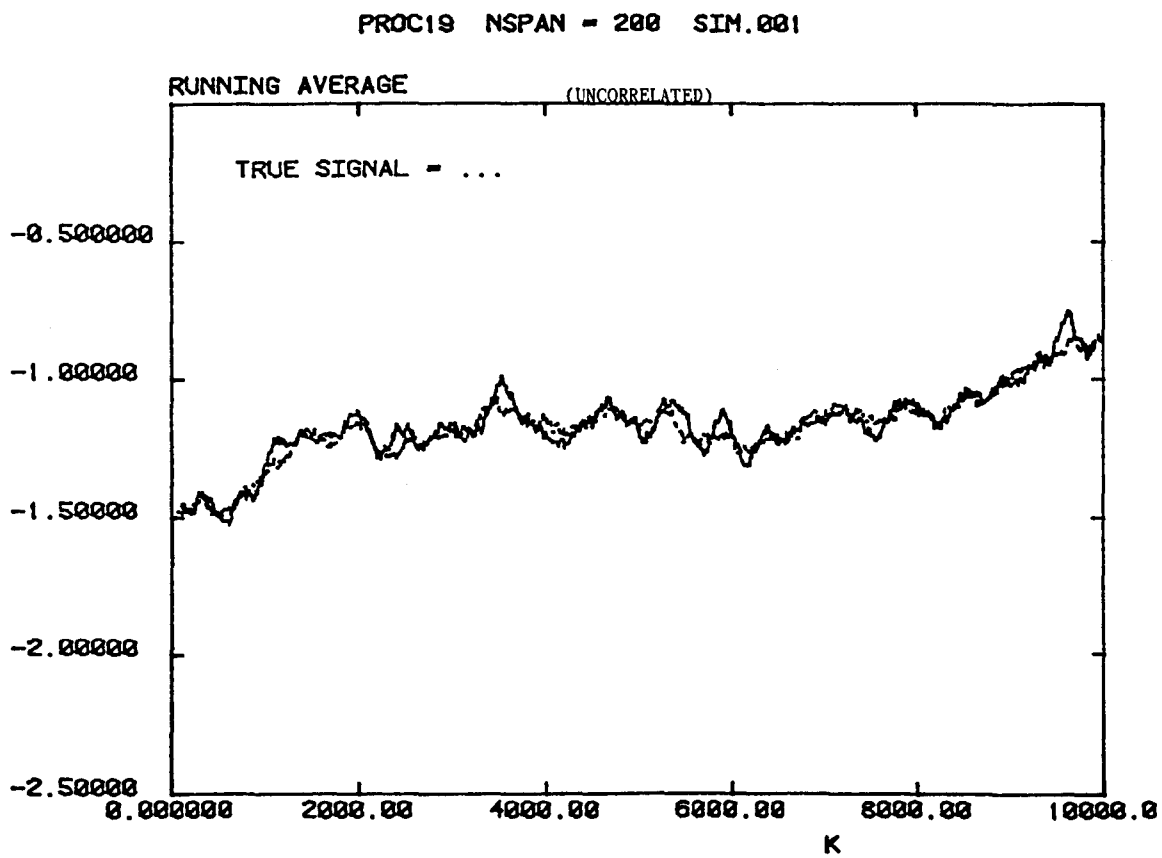


FIGURE 7.7 (f)



Running Average Estimates of Simulated Data: Uncorrelated Speckle (above)
: Correlated Speckle (below)

FIGURE 7.7 (g)

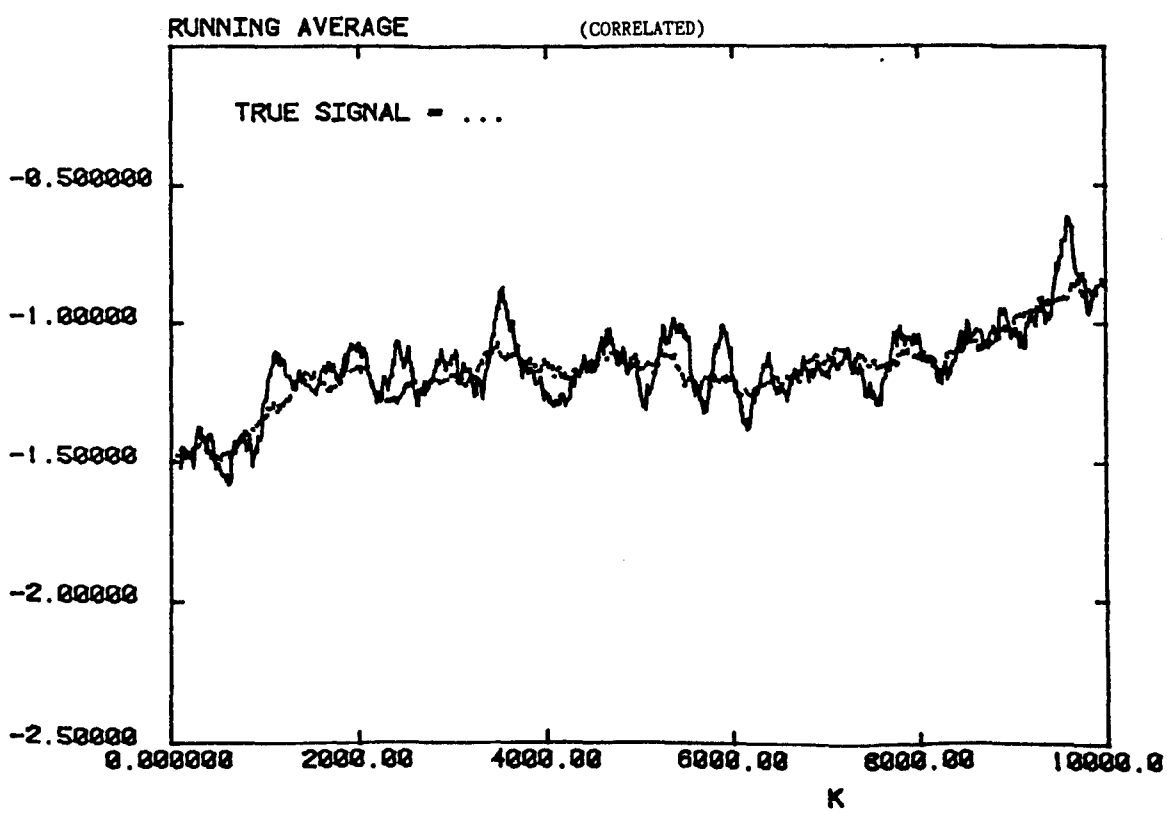
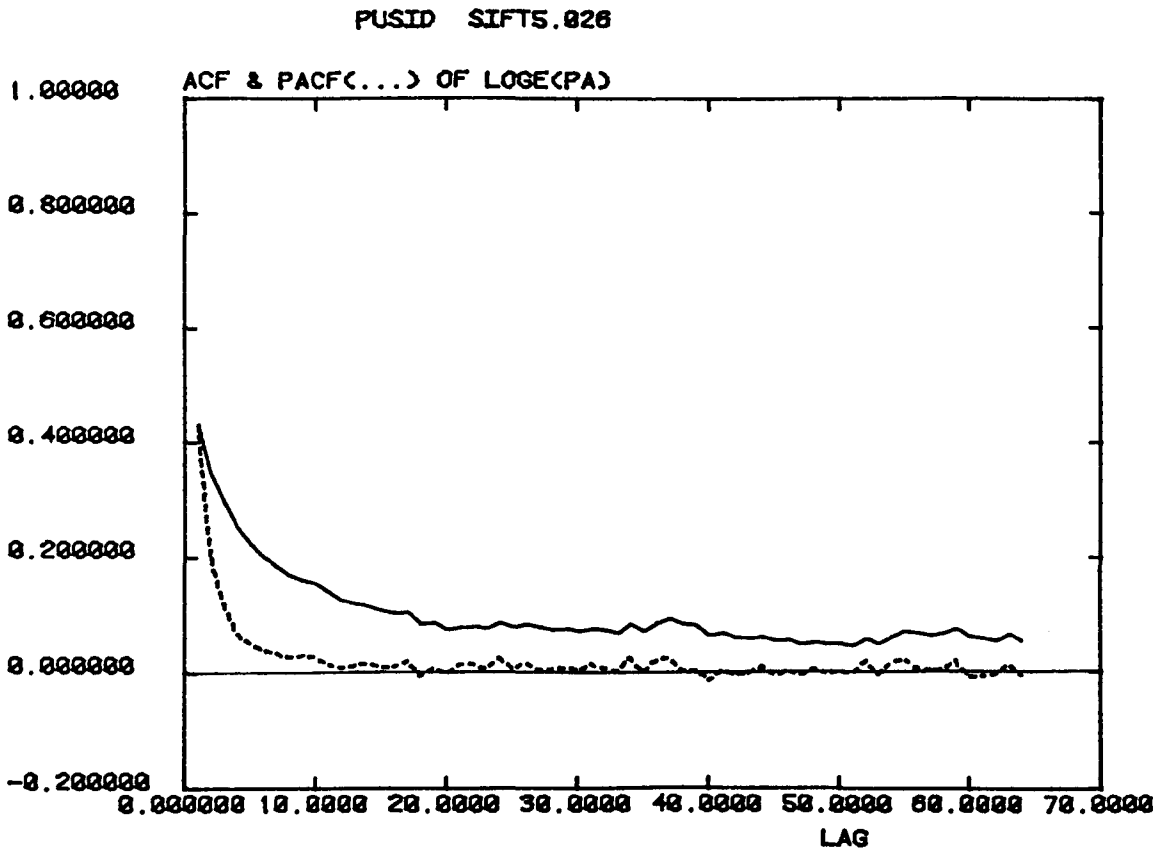


FIGURE 7.8 (a)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT 5·026 : Channel A (above)
: Channel B (below)

FIGURE 7.8 (b)

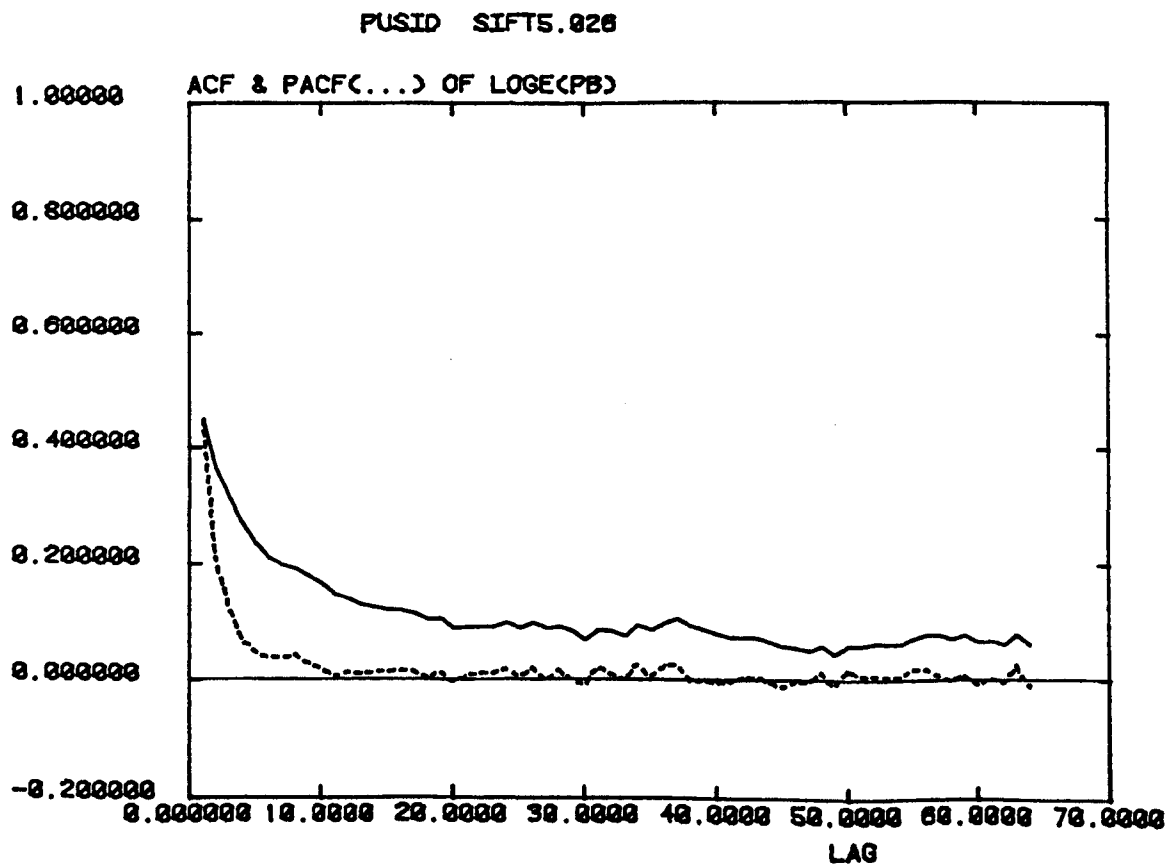
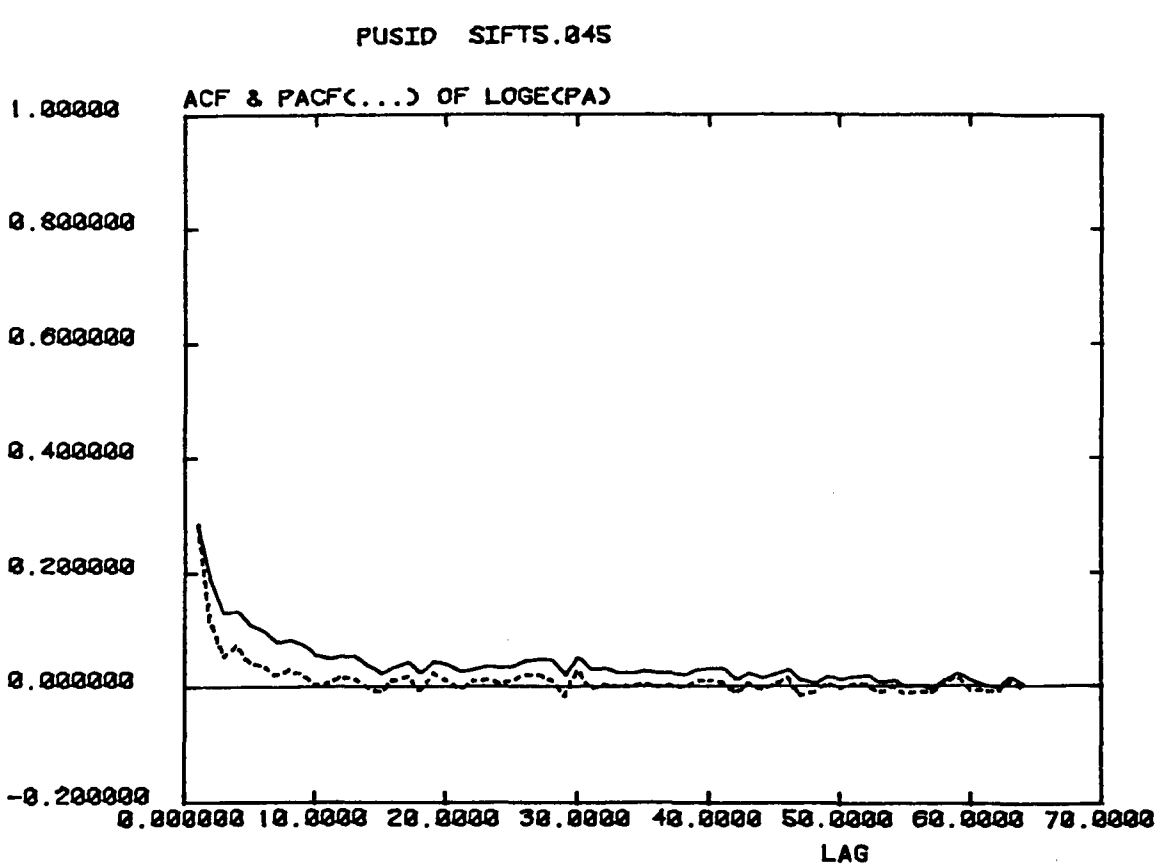
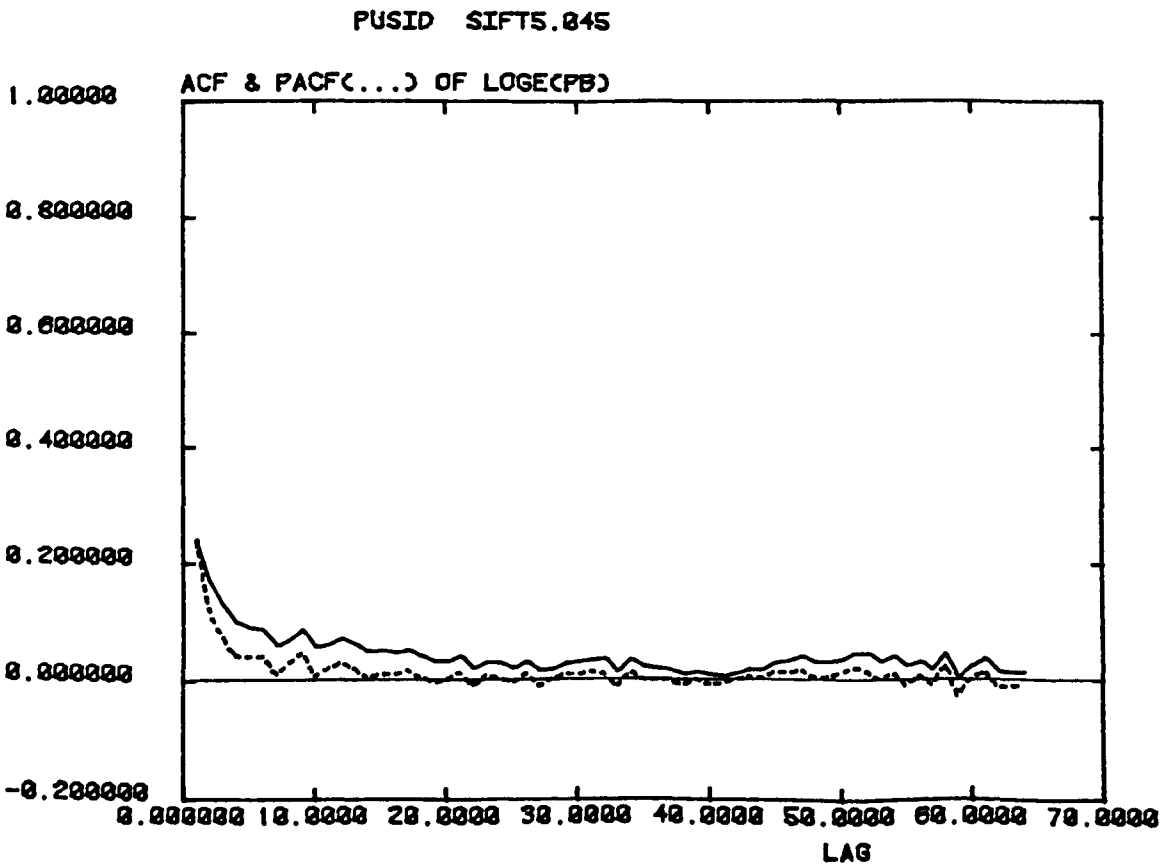


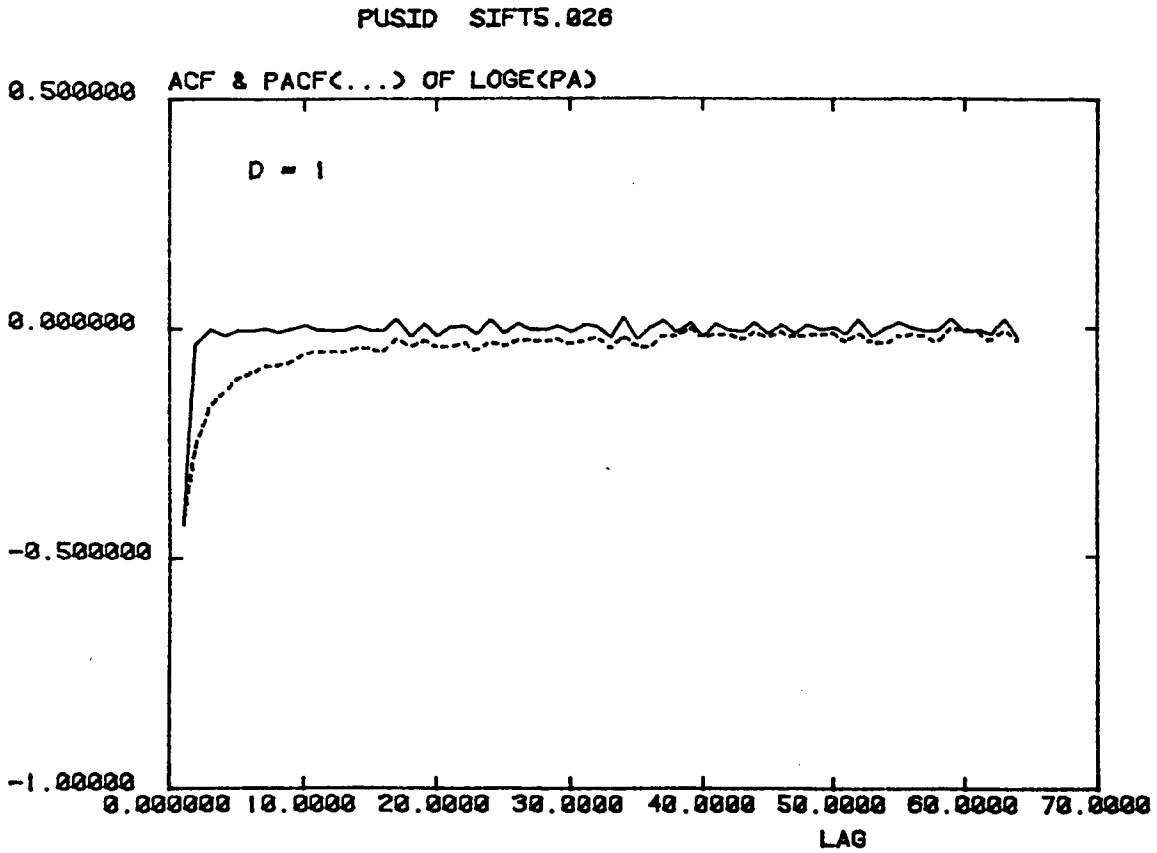
FIGURE 7.8 (c)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT5·045 : Channel A (above)
: Channel B (below)

FIGURE 7.8 (d)





acf and pacf for the First Order Difference of the Log_e (Normalised Power) Data of SIFT5·045: Channel A (above)
Channel B (below)

FIGURE 7.8 (f)

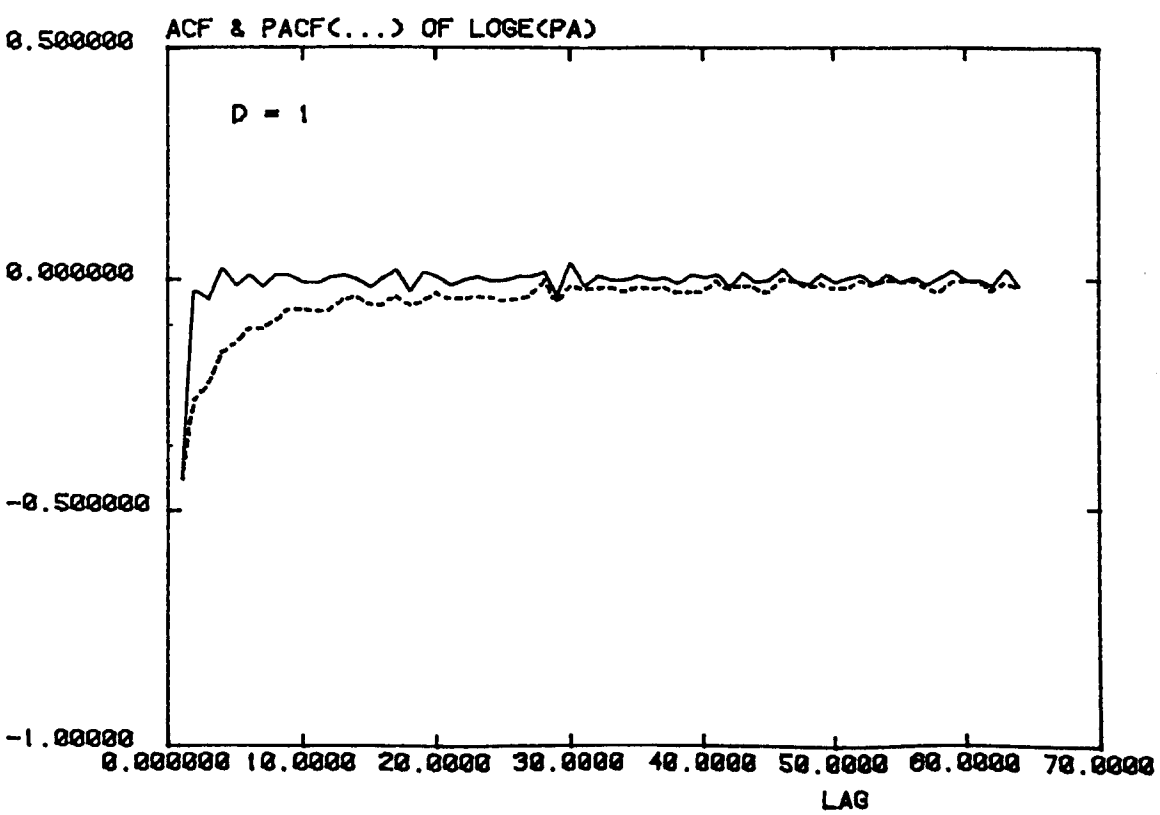
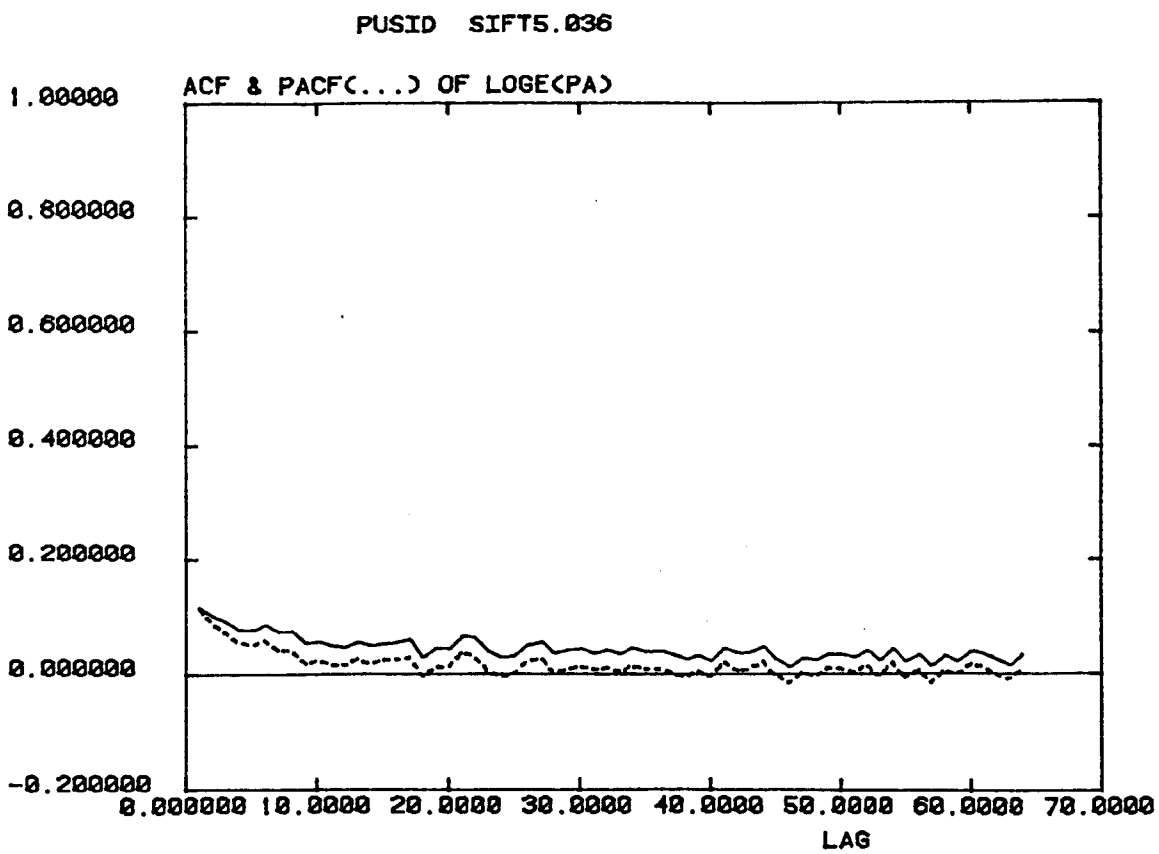


FIGURE 7.8 (g)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT 5·036 : Channel A (above)
: Channel B (below)

FIGURE 7.8 (h)

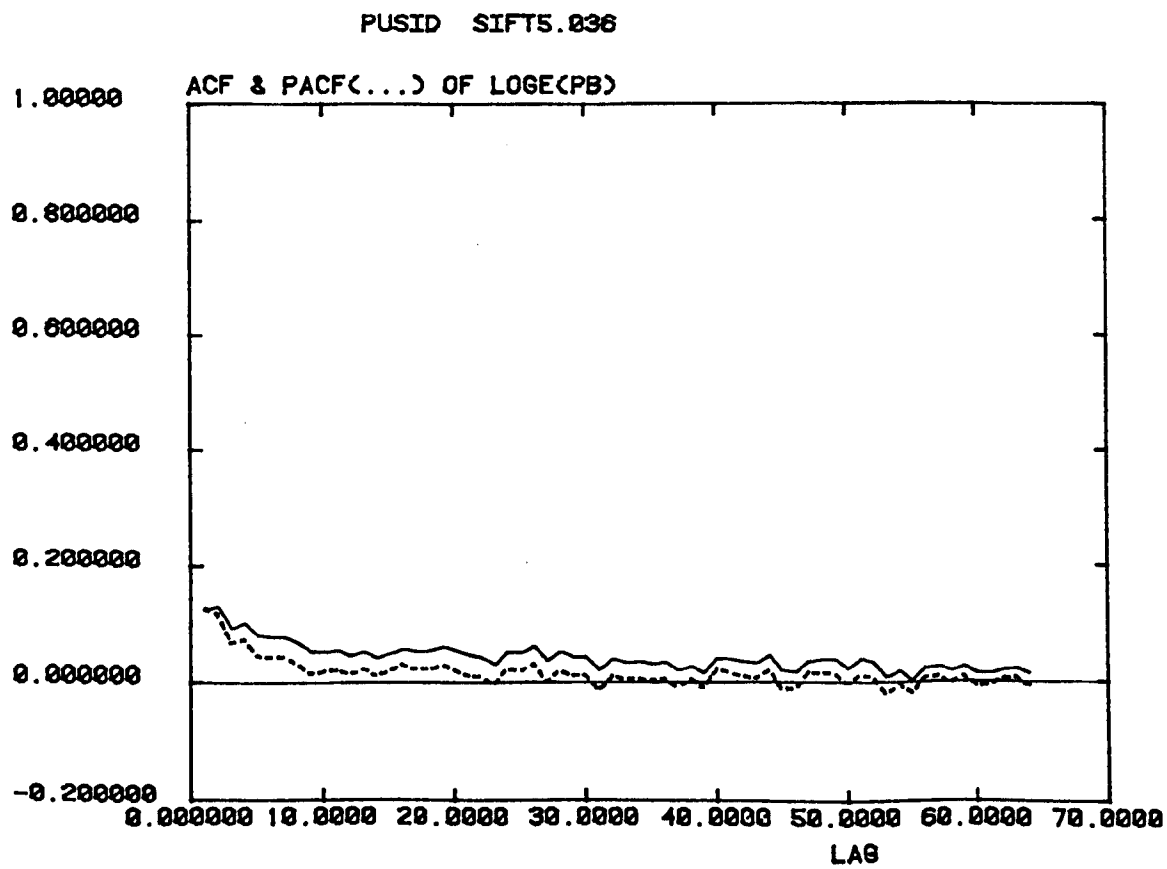
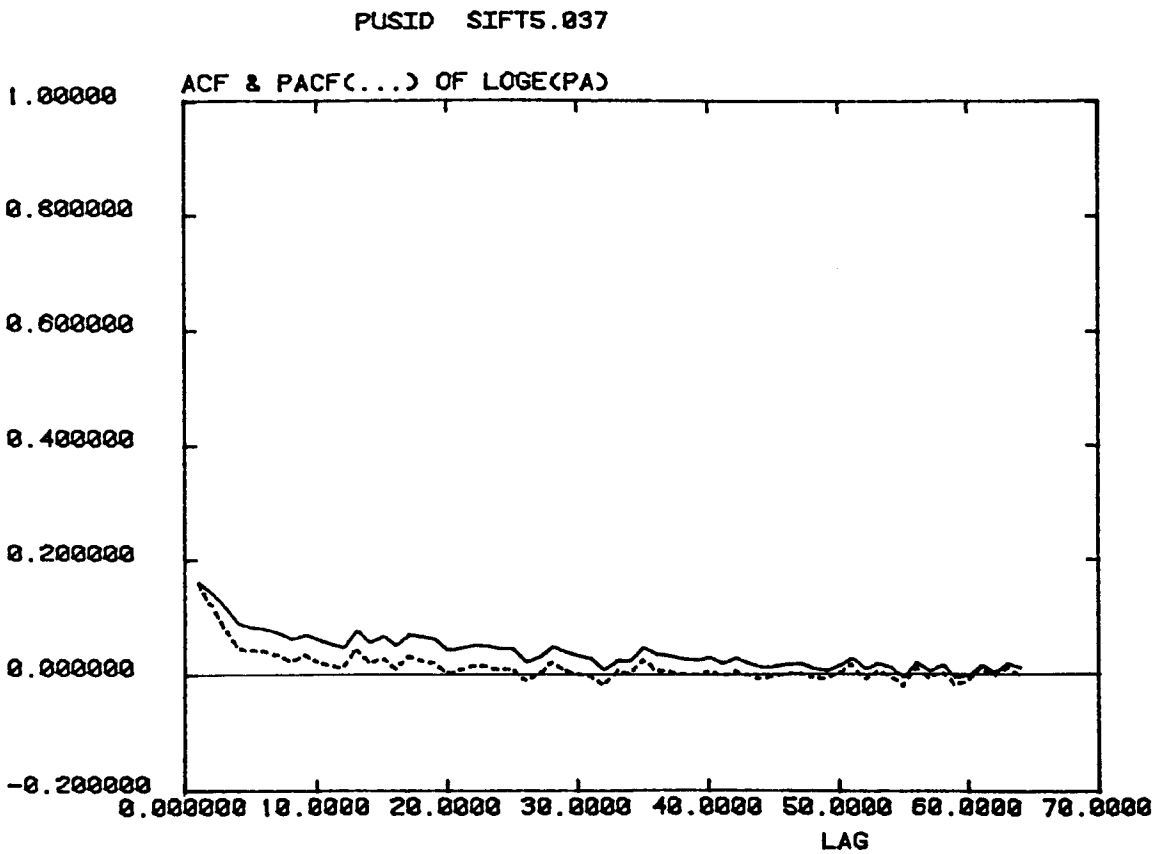


FIGURE 7.8 (i)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT 5·037 : Channel A (above)
: Channel B (below)

FIGURE 7.8 (j)

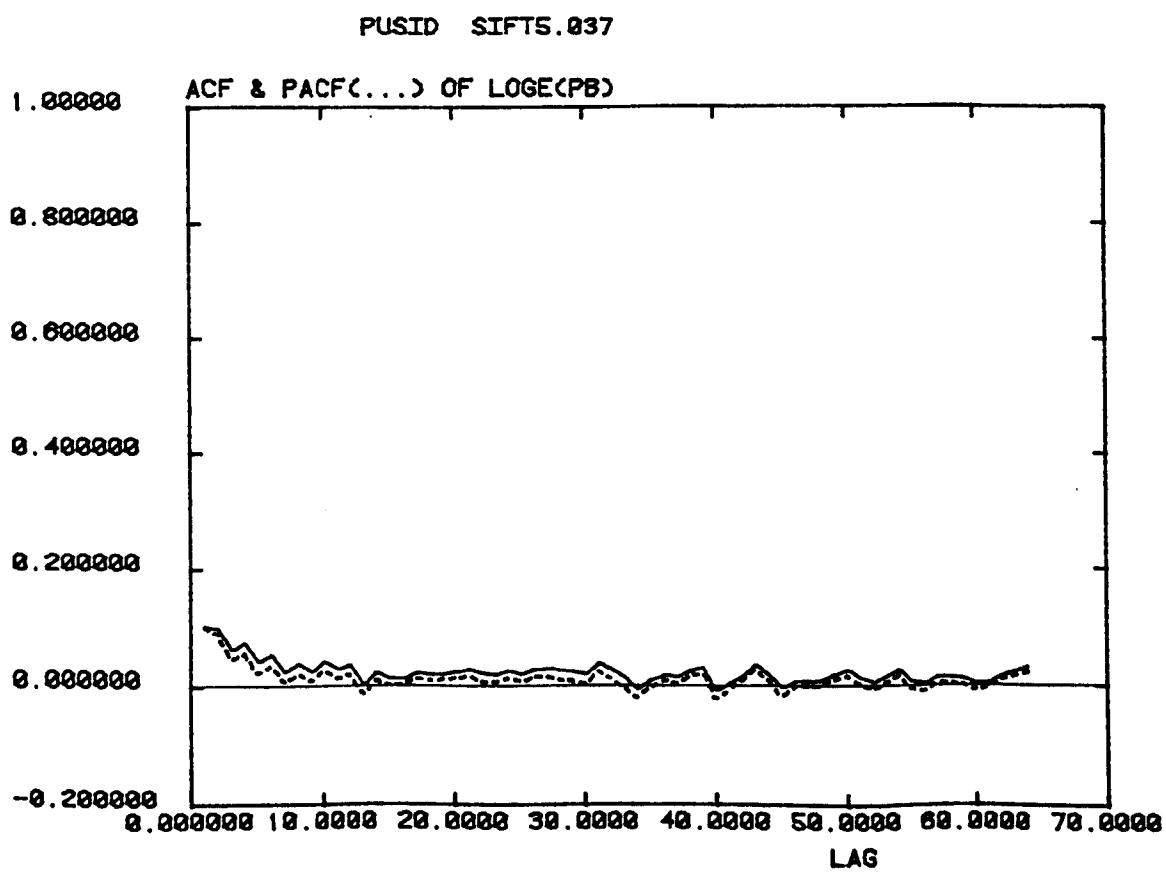
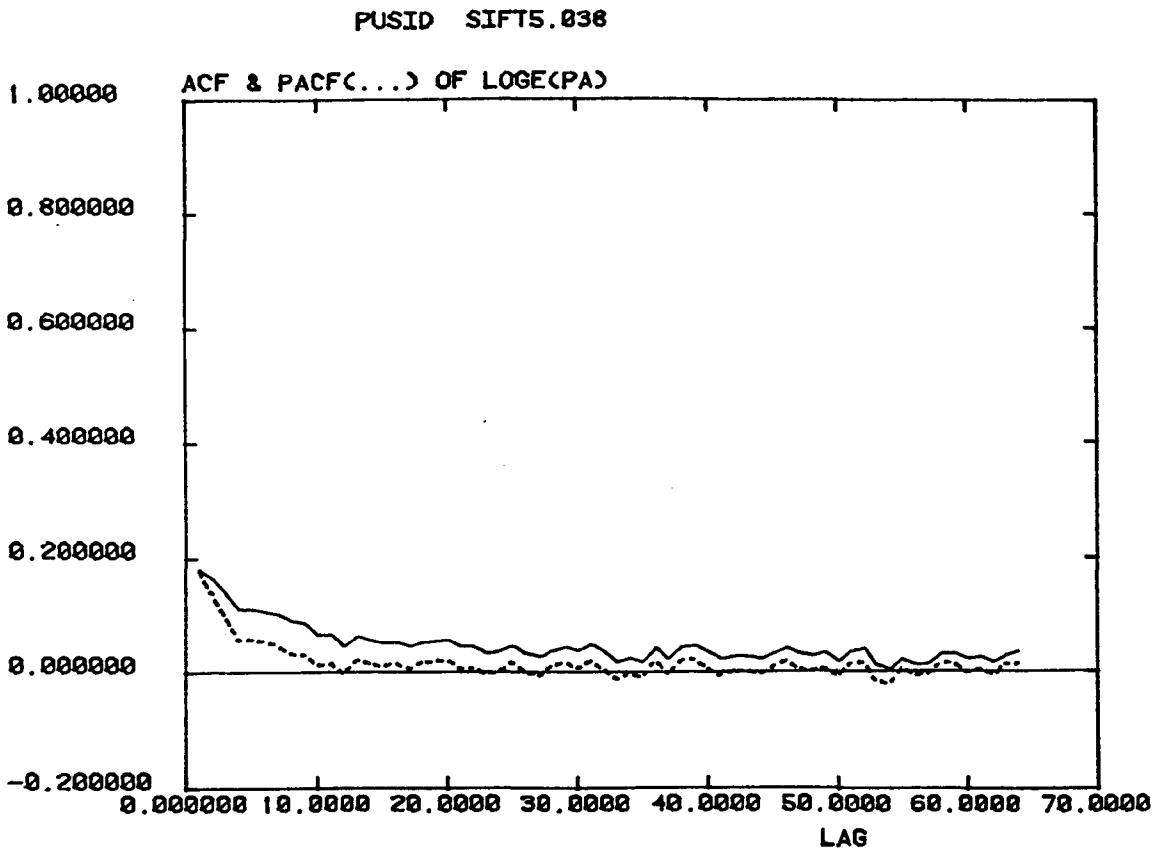


FIGURE 7.8 (k)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT 5.038 : Channel A (above)
: Channel B (below)

FIGURE 7.8 (l)

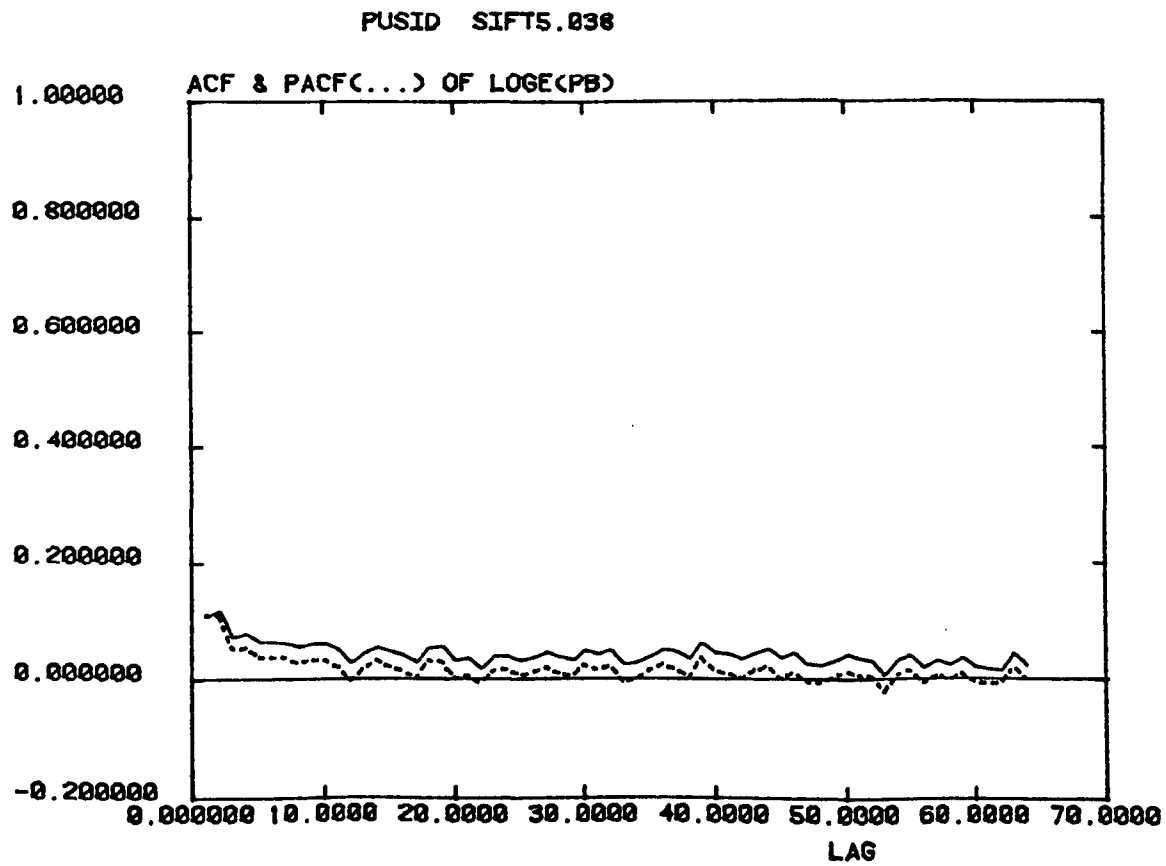
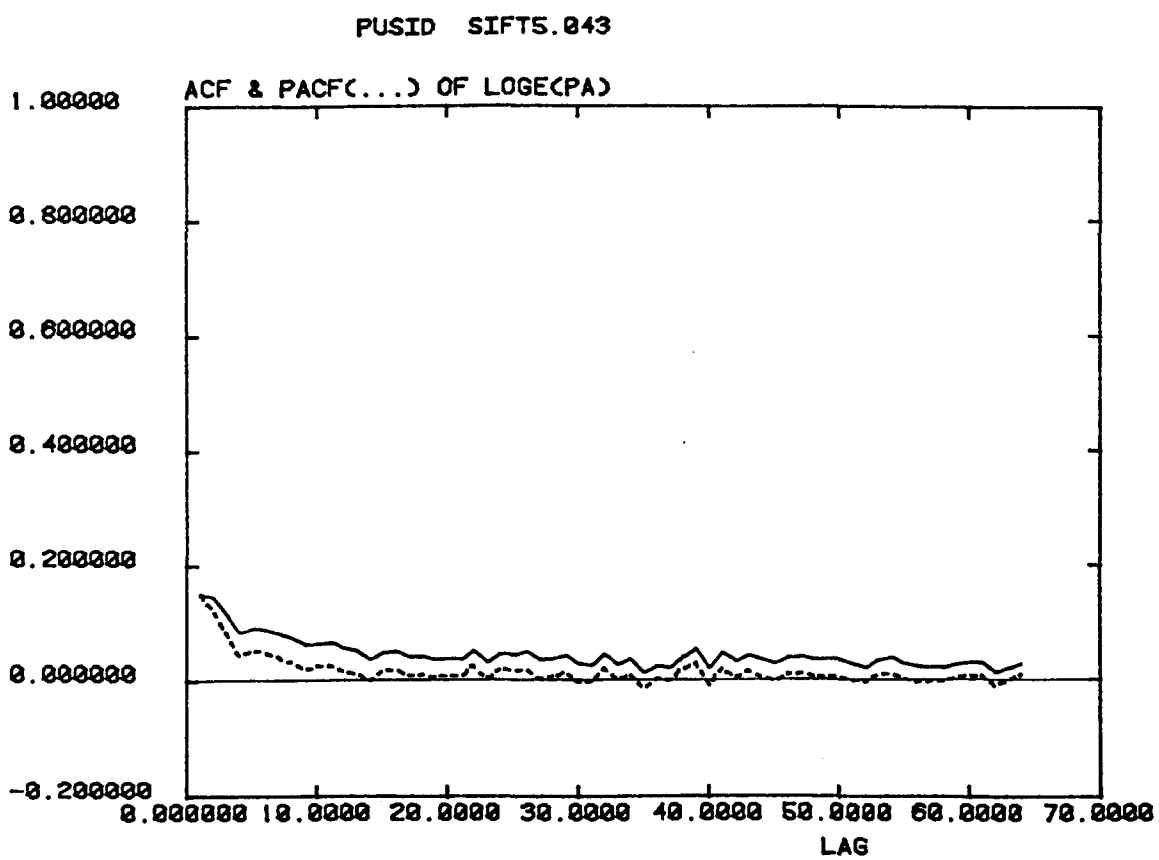


FIGURE 7.8 (m)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT 5·043 : Channel A (above)
: Channel B (below)

FIGURE 7.8 (n)

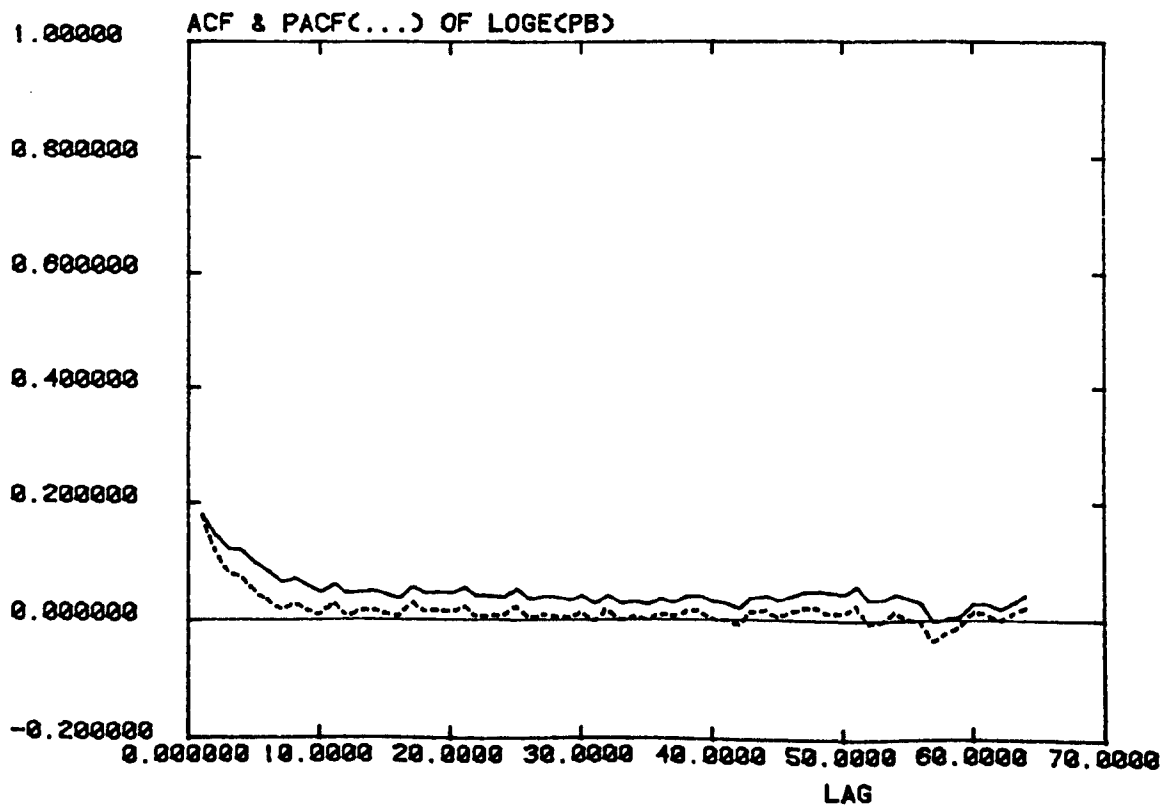
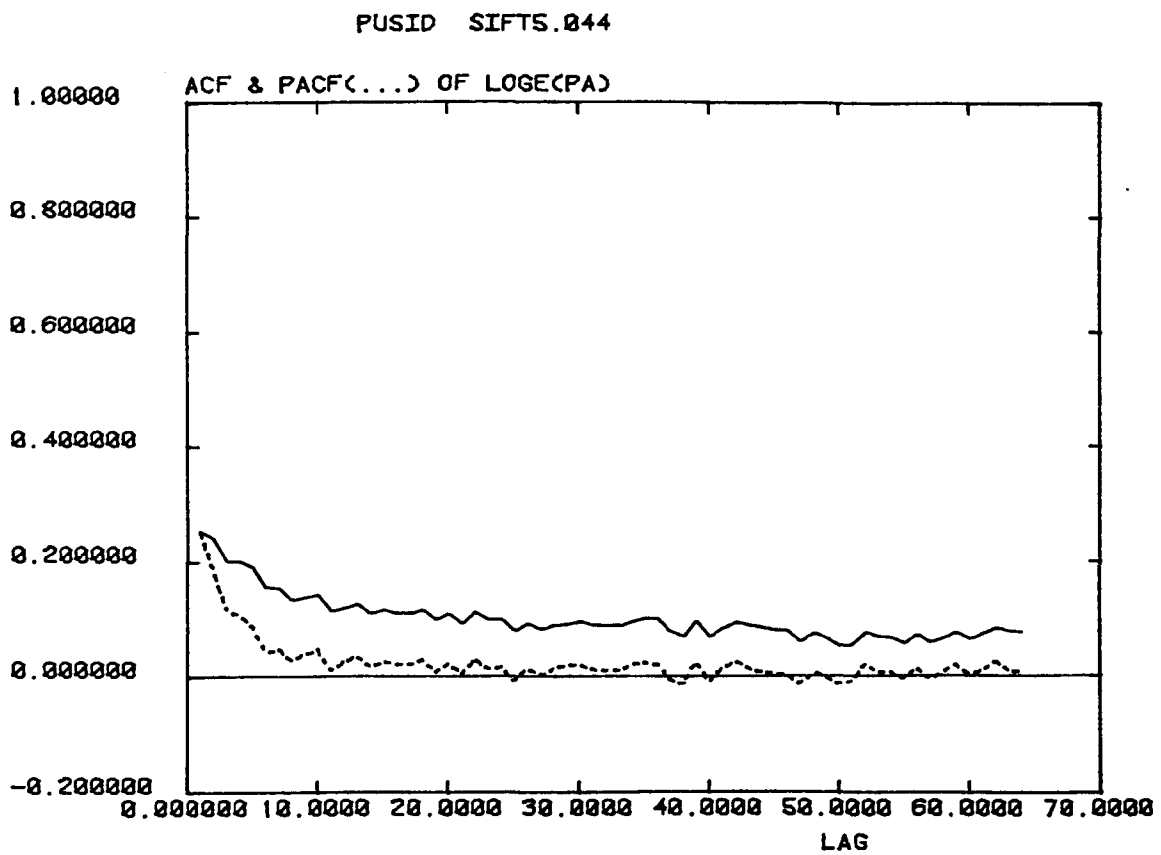


FIGURE 7.8 (o)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT 5.044 : Channel A (above)
: Channel B (below)

FIGURE 7.8 (p)

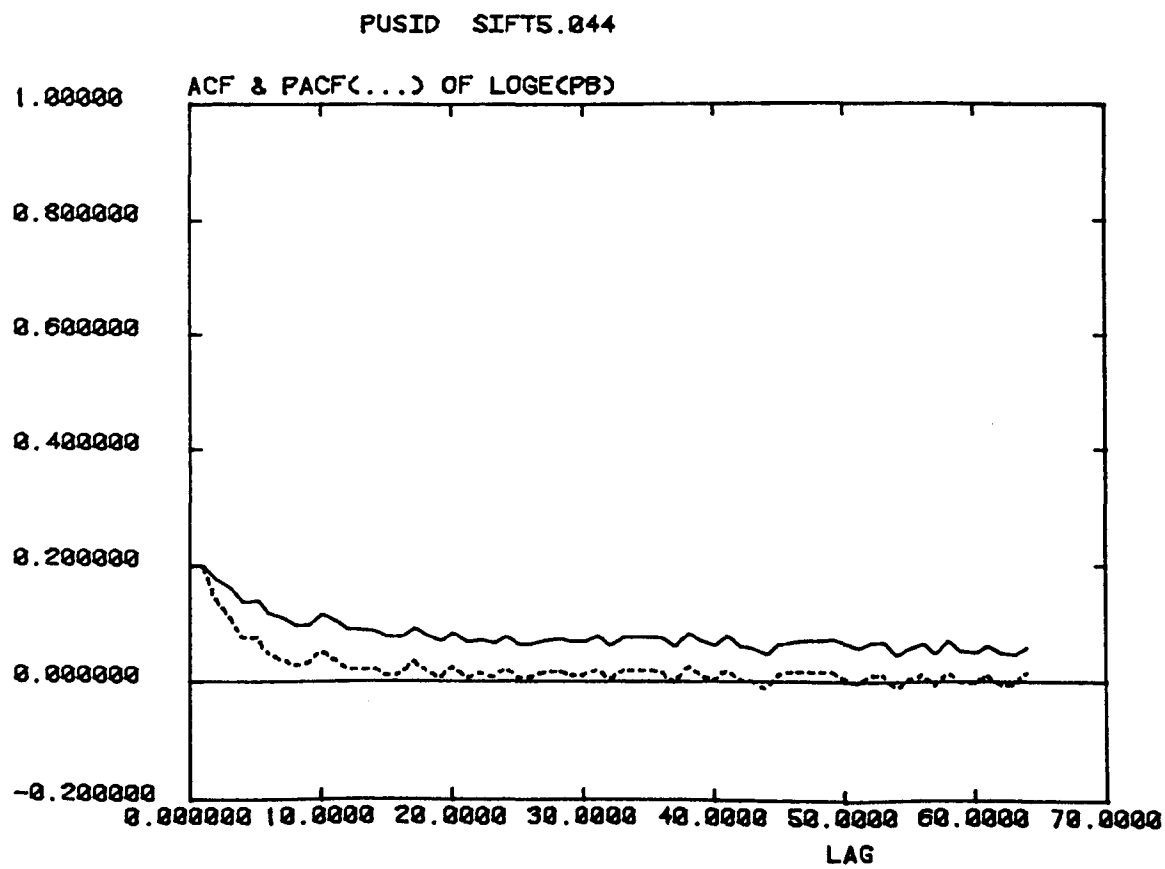
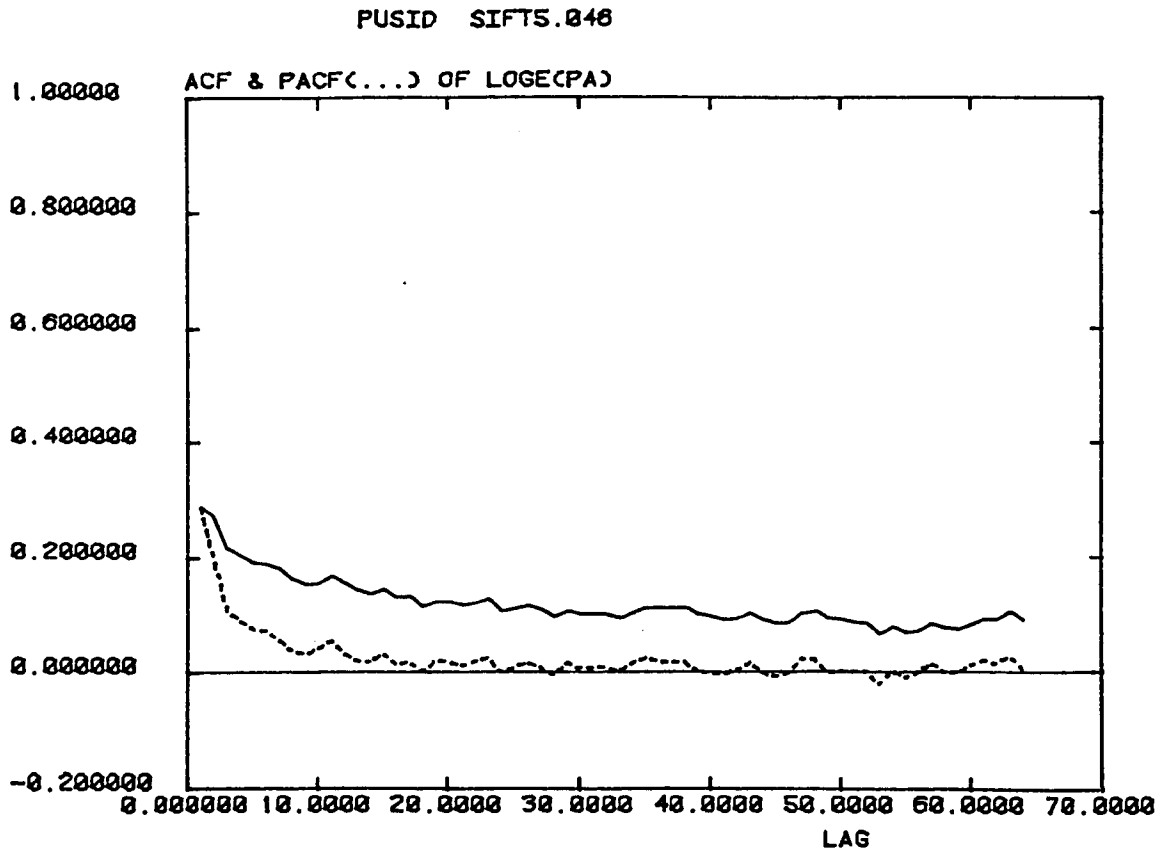


FIGURE 7.8 (q)



acf and pacf of the $\log_e$ (Normalised Power)
Data for SIFT 5·046 : Channel A (Above)
: Channel B (Below)

FIGURE 7.8 (r)

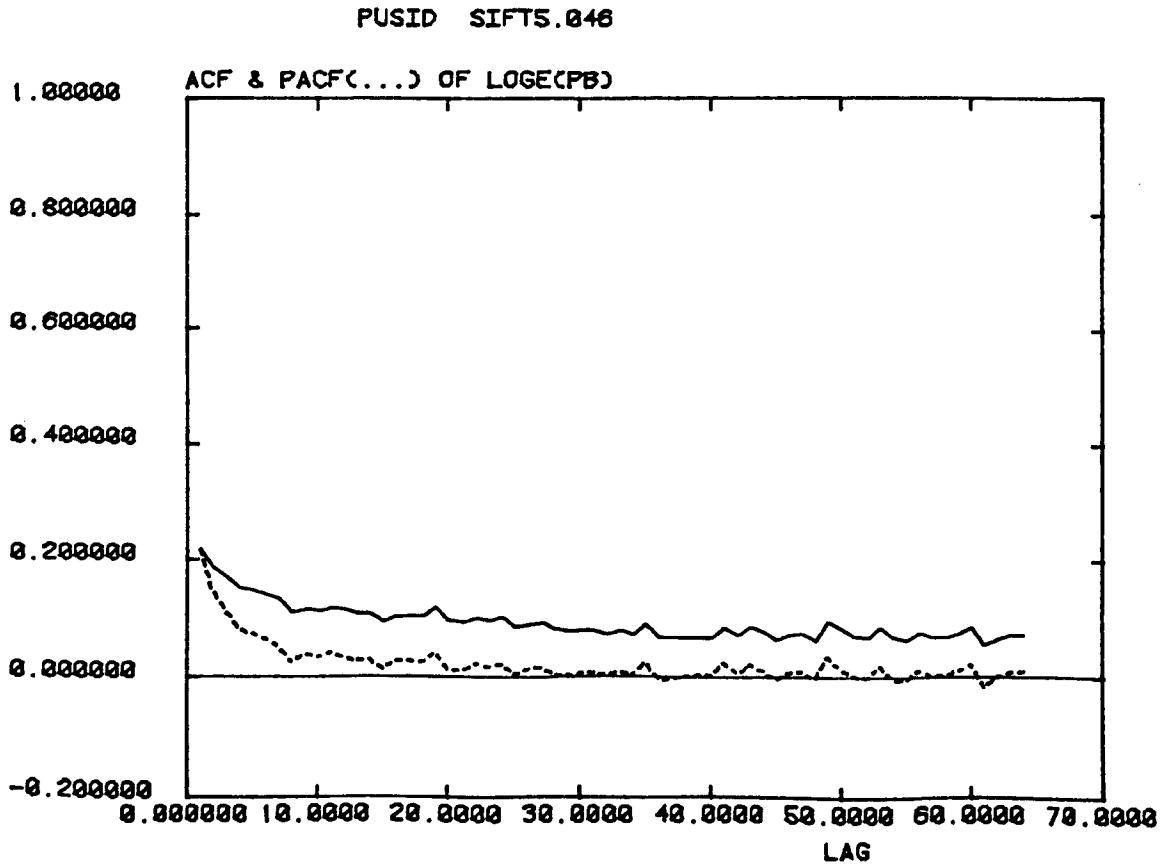


FIGURE 7.9

Acf and *Pacf* of Autoregressive Models.

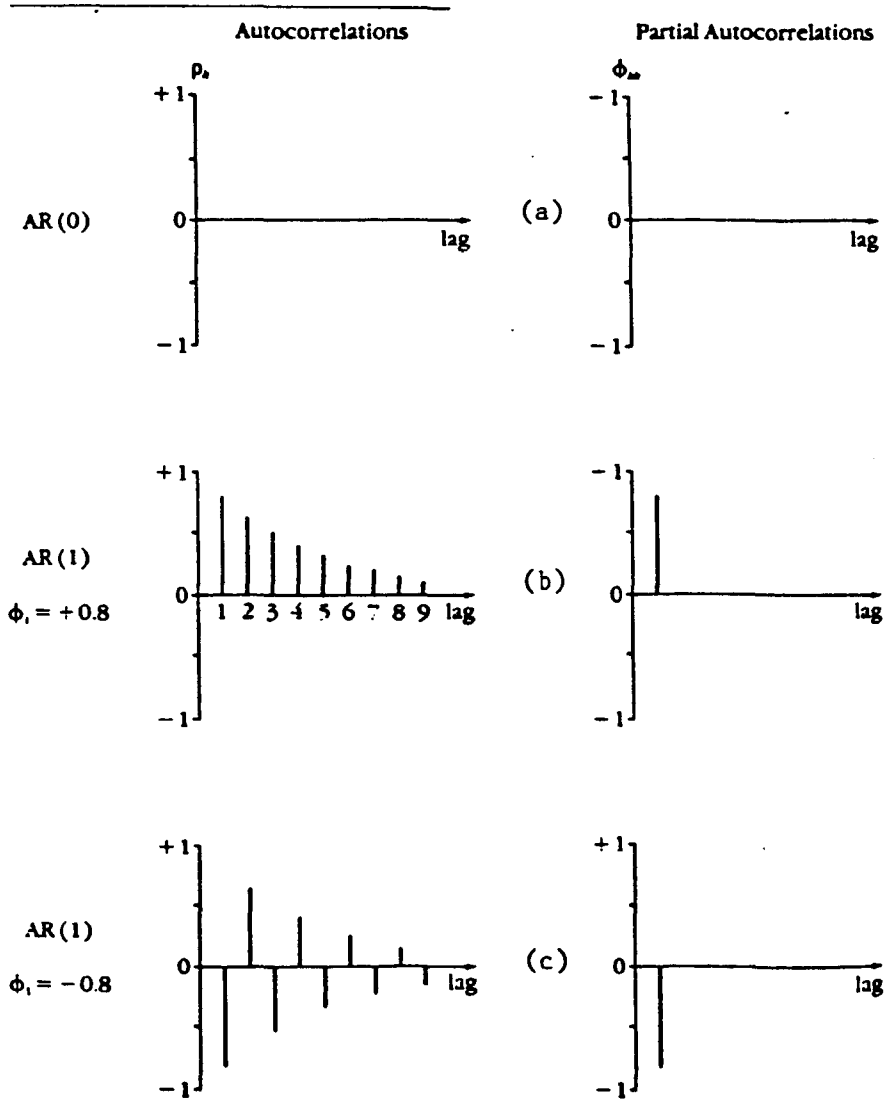


FIGURE 7.9

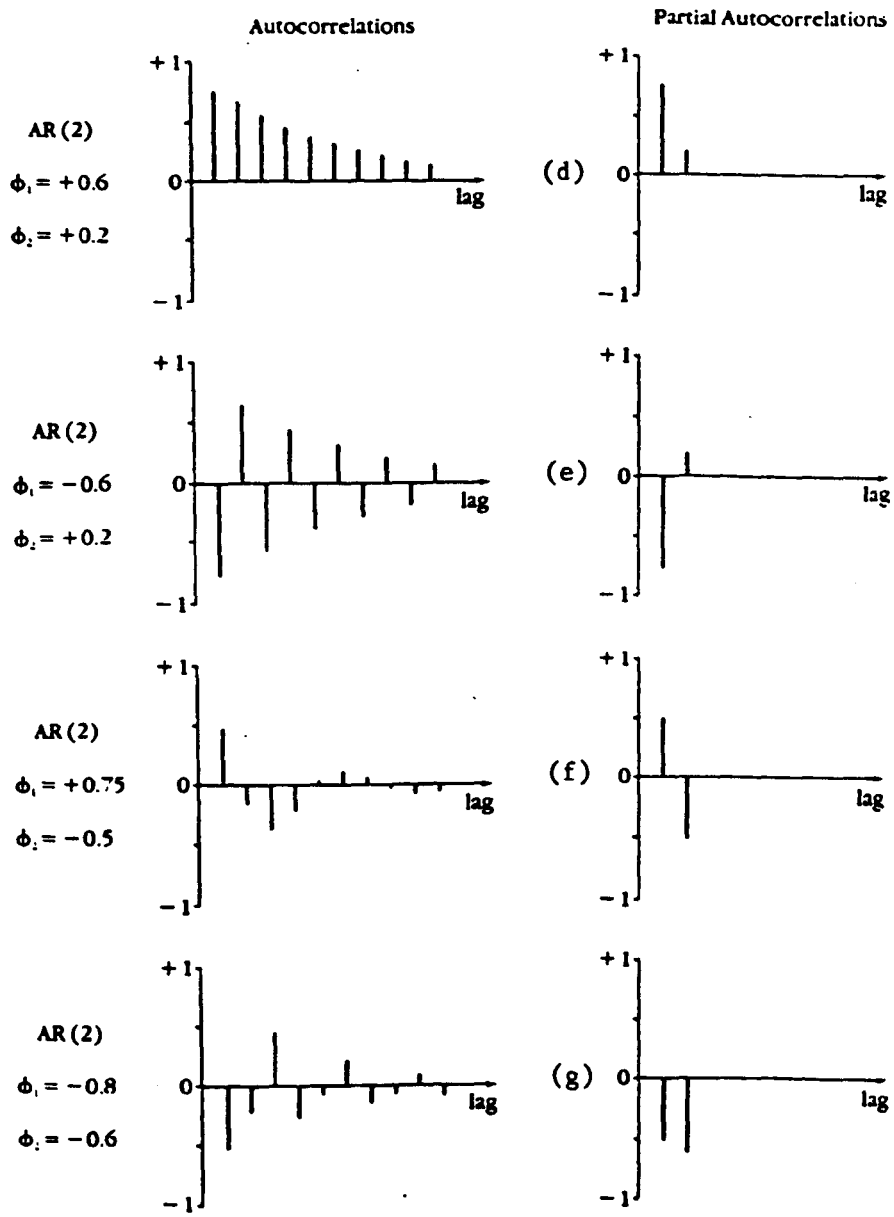

Autocorrelations
Partial Autocorrelations
AR (2)
$\phi_1 = +0.6$
$\phi_2 = +0.2$
(d)
AR (2)
$\phi_1 = -0.6$
$\phi_2 = +0.2$
(e)
AR (2)
$\phi_1 = +0.75$
$\phi_2 = -0.5$
(f)
AR (2)
$\phi_1 = -0.8$
$\phi_2 = -0.6$
(g)
+1
0
-1
lag

FIGURE 7.9

Acf and *Pacf* of Moving Average Models.

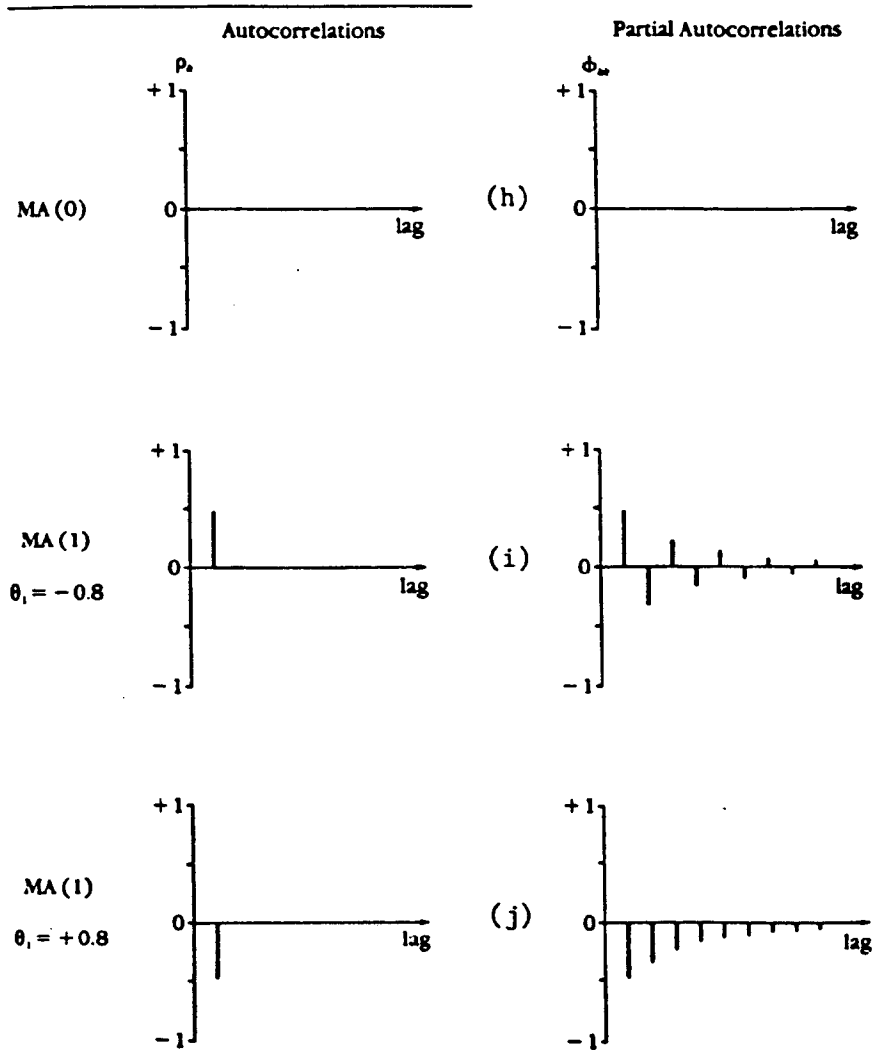


FIGURE 7.9

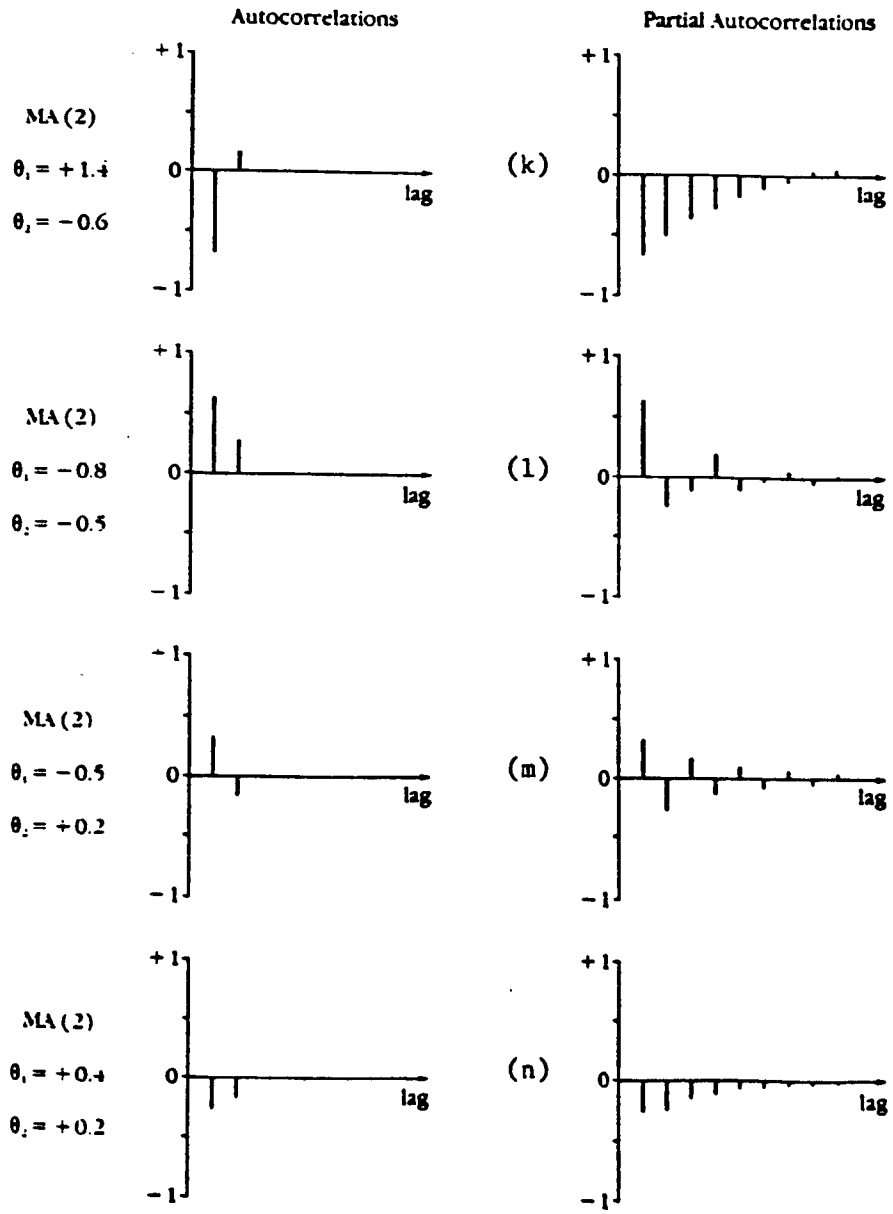
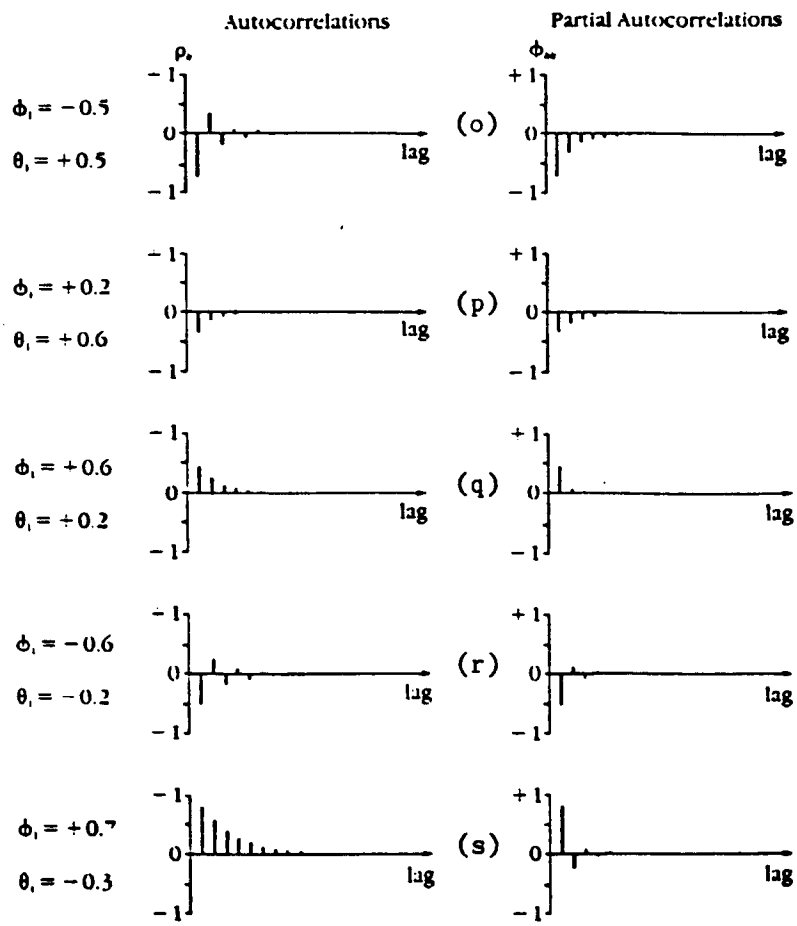

Autocorrelations
Partial Autocorrelations
MA (2)
θ1 = +1.4
θ2 = −0.6
(k)
MA (2)
θ1 = −0.8
θ2 = −0.5
(l)
MA (2)
θ1 = −0.5
θ2 = +0.2
(m)
MA (2)
θ1 = +0.4
θ2 = +0.2
(n)
+1
0
−1
lag

FIGURE 7.9

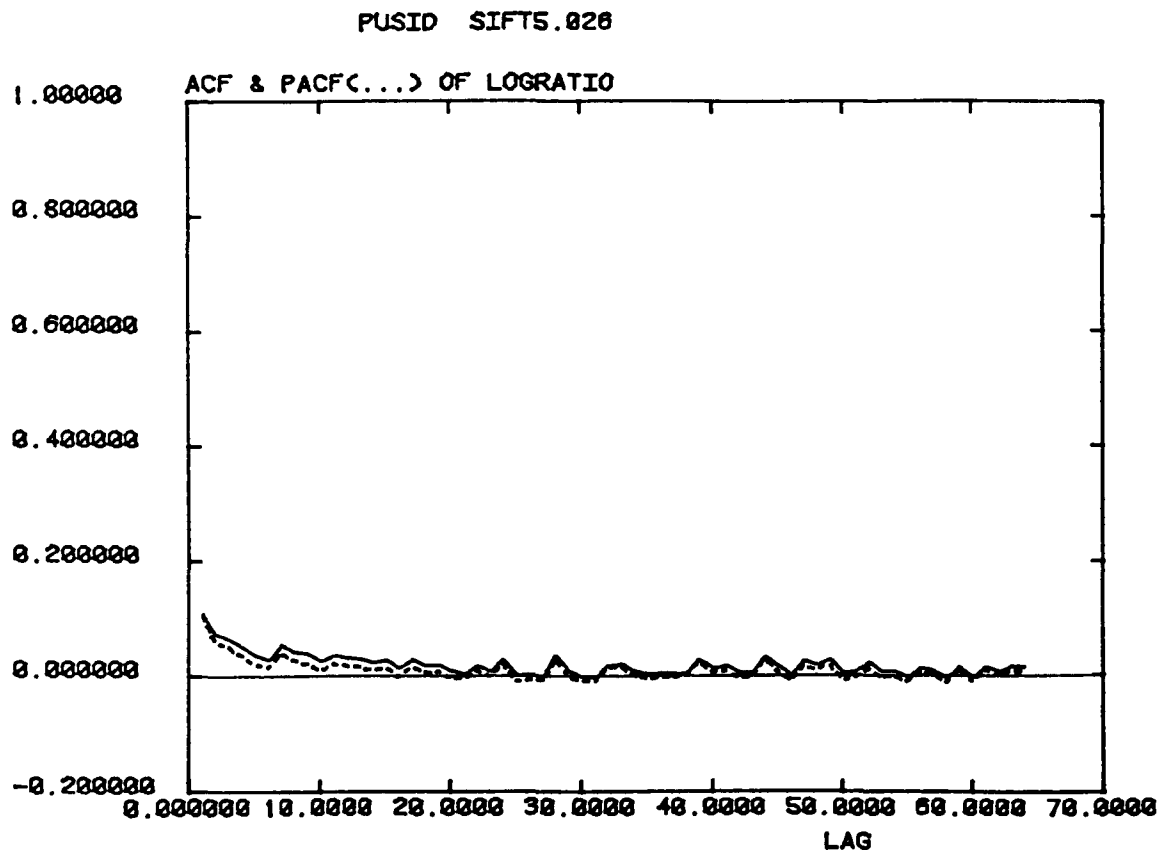


Acf and *Pacf* of ARMA(1,1) Models.

7.9 (t) Autocorrelation and Partial Autocorrelation Functions of Nonseasonal Models

MODEL	AUTOCORRELATION FUNCTION	PARTIAL AUTOCORRELATION FUNCTION
AR(p)	Tails off	Cuts off after lag p
MA(q)	Cuts off after lag q	Tails off
ARMA(p,q)	Tails off	Tails off

FIGURE 7.10 (a)



acf and pact of $\log_e$ Ratio Data: SIFT 5·026 (above)

: SIFT 5·045 (below)

FIGURE 7.10 (b)

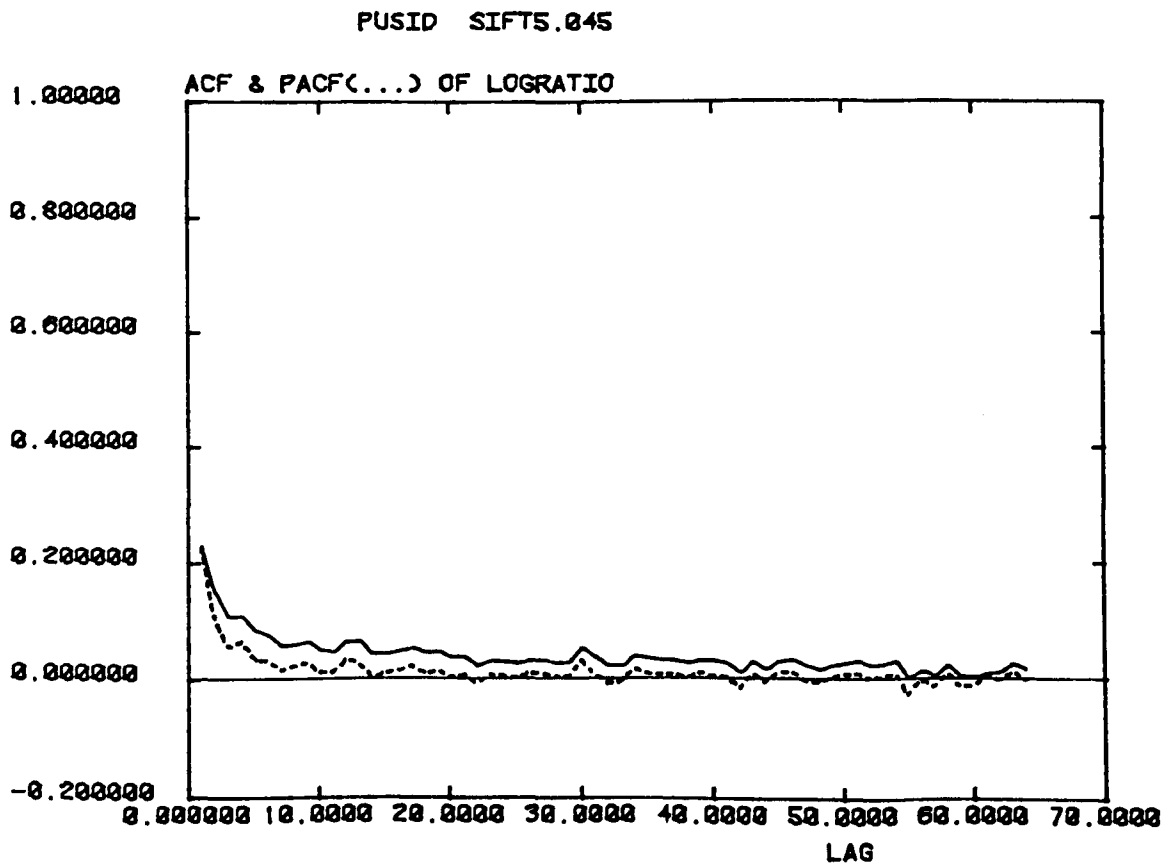
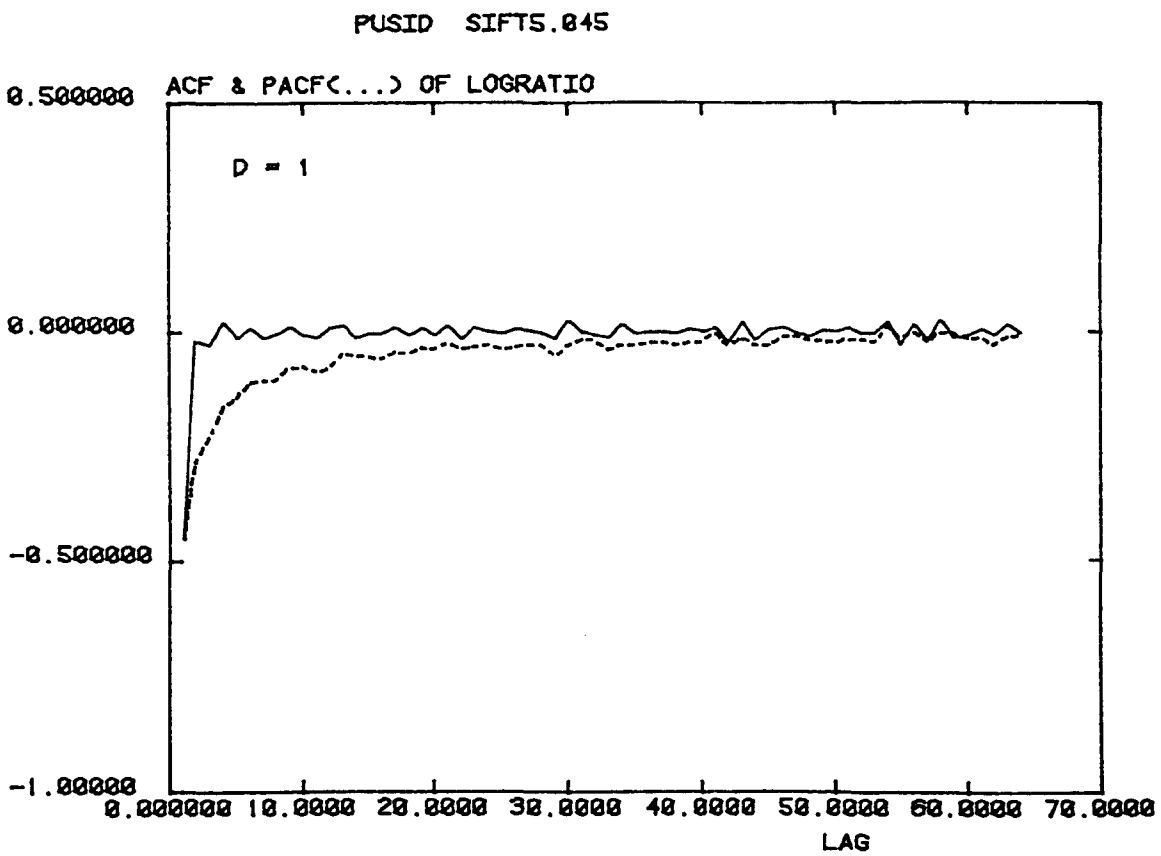


FIGURE 7.10 (c)



acf and pacf of Differenced Log Ratio Data for SIFT 5·045 (above)
acf and pacf of Log Ratio Data for SIFT 5·036 (below)

FIGURE 7.10 (d)

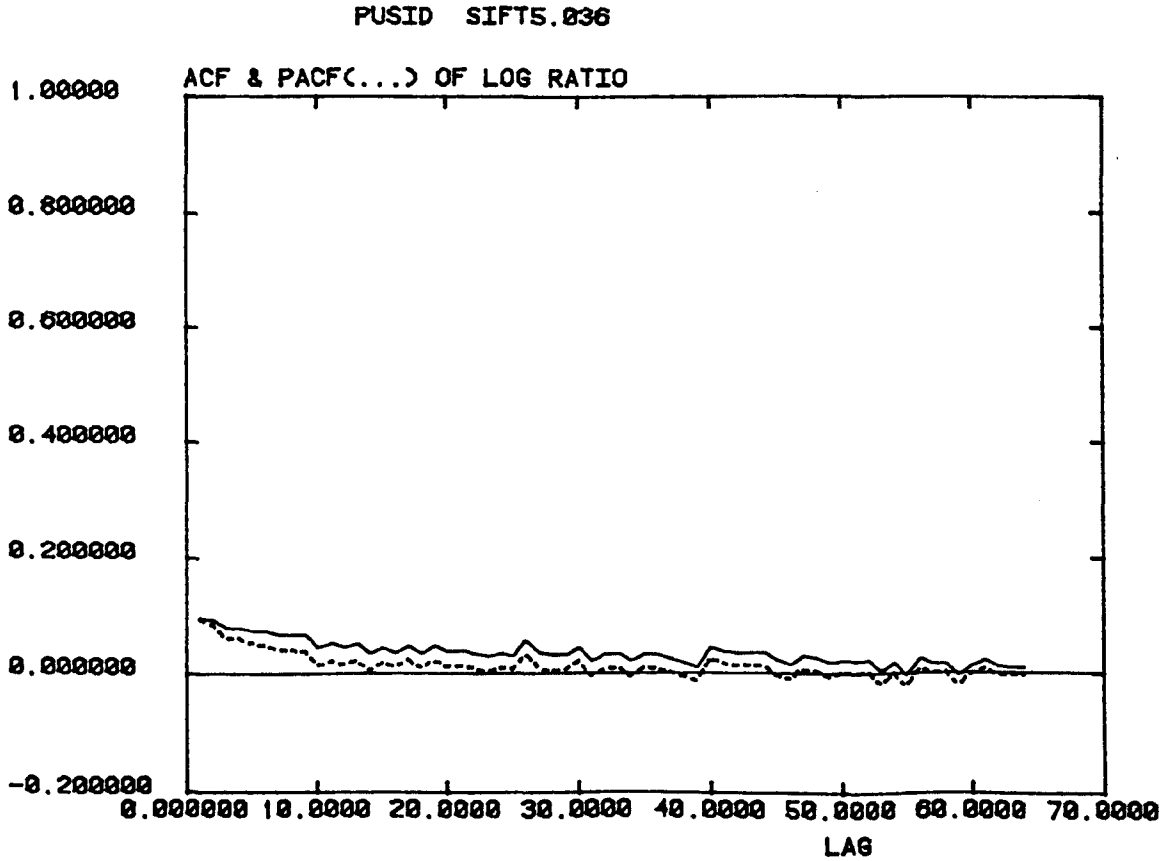
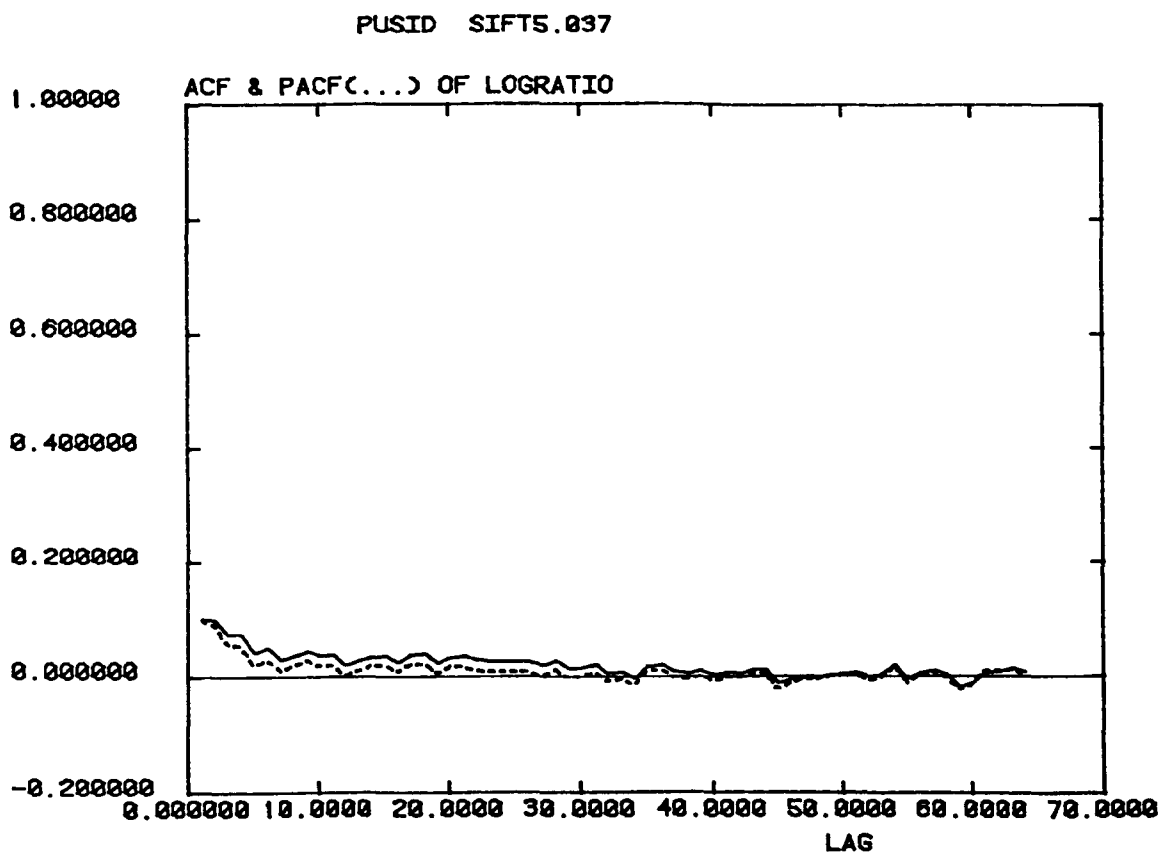


FIGURE 7.10 (e)



acf and pacf of Log Ratio Data: SIFT5·037 (above)

: SIFT5·038 (below)

FIGURE 7.10 (f)

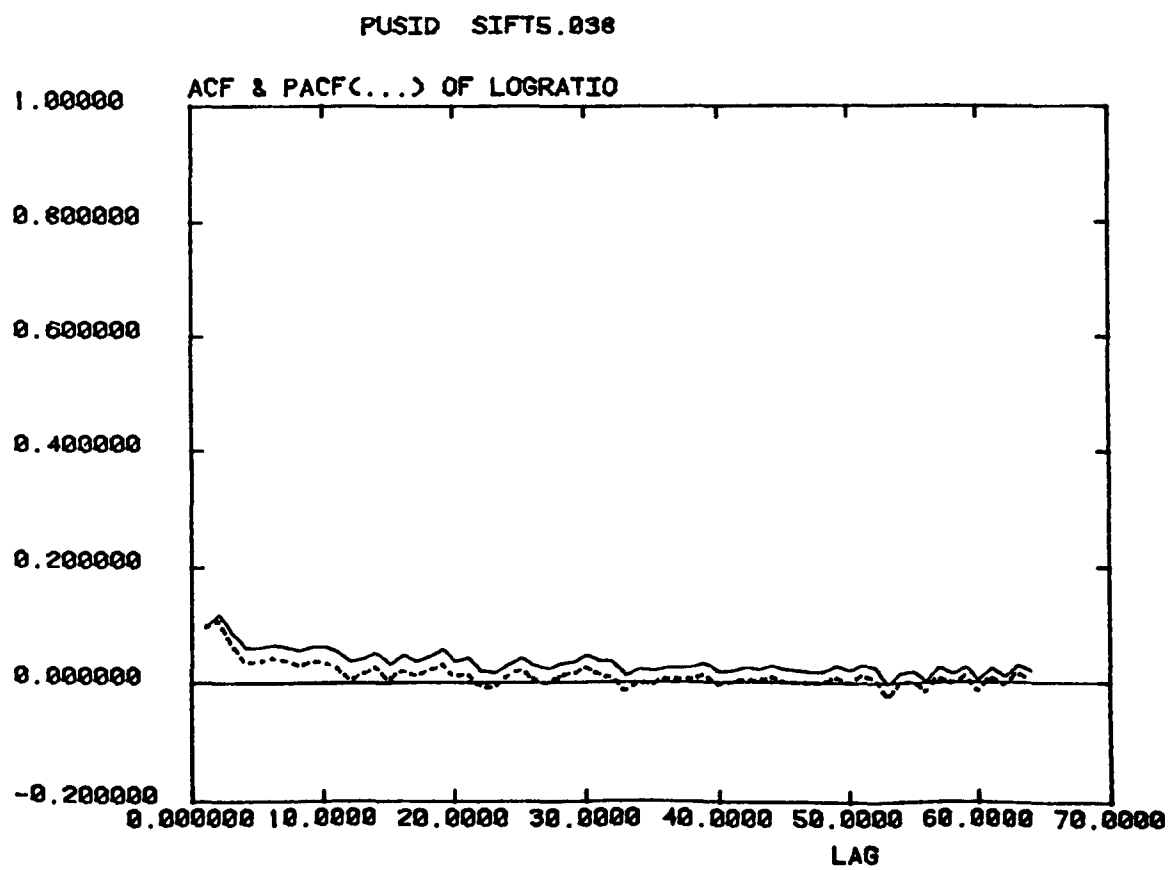
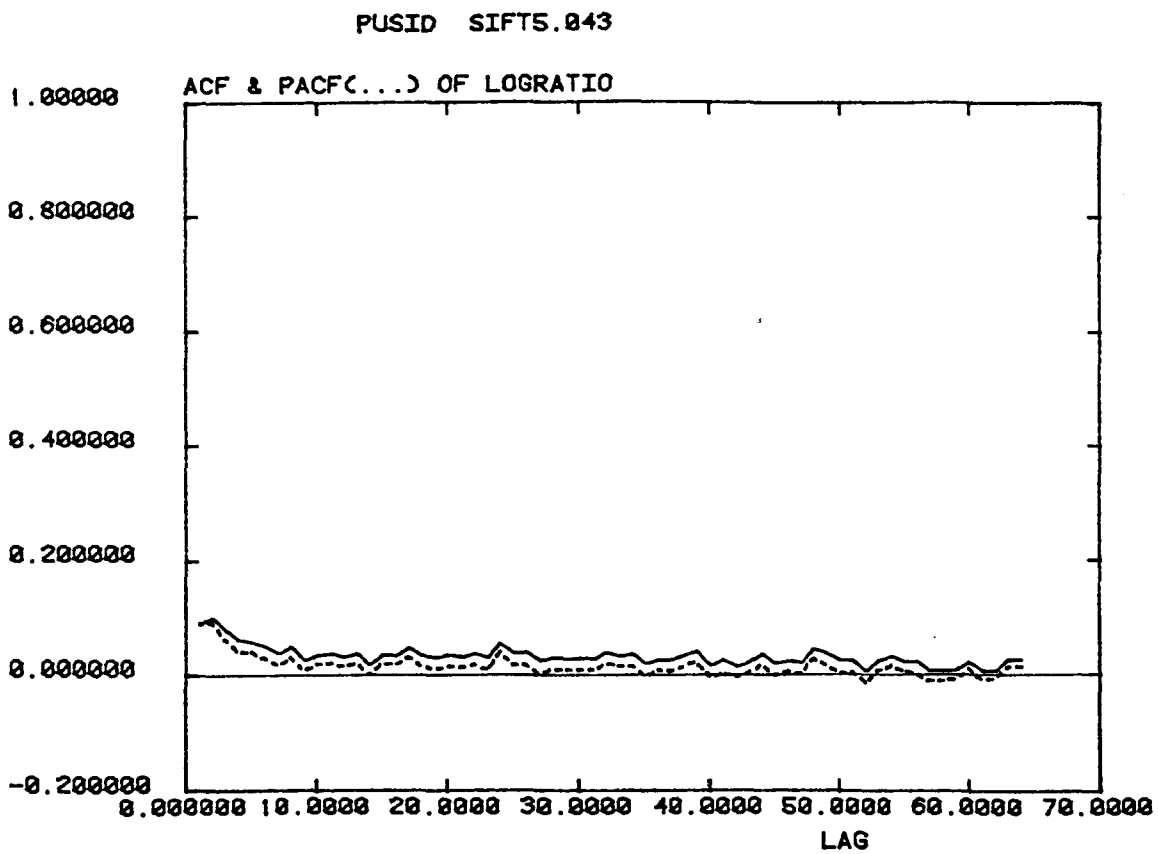
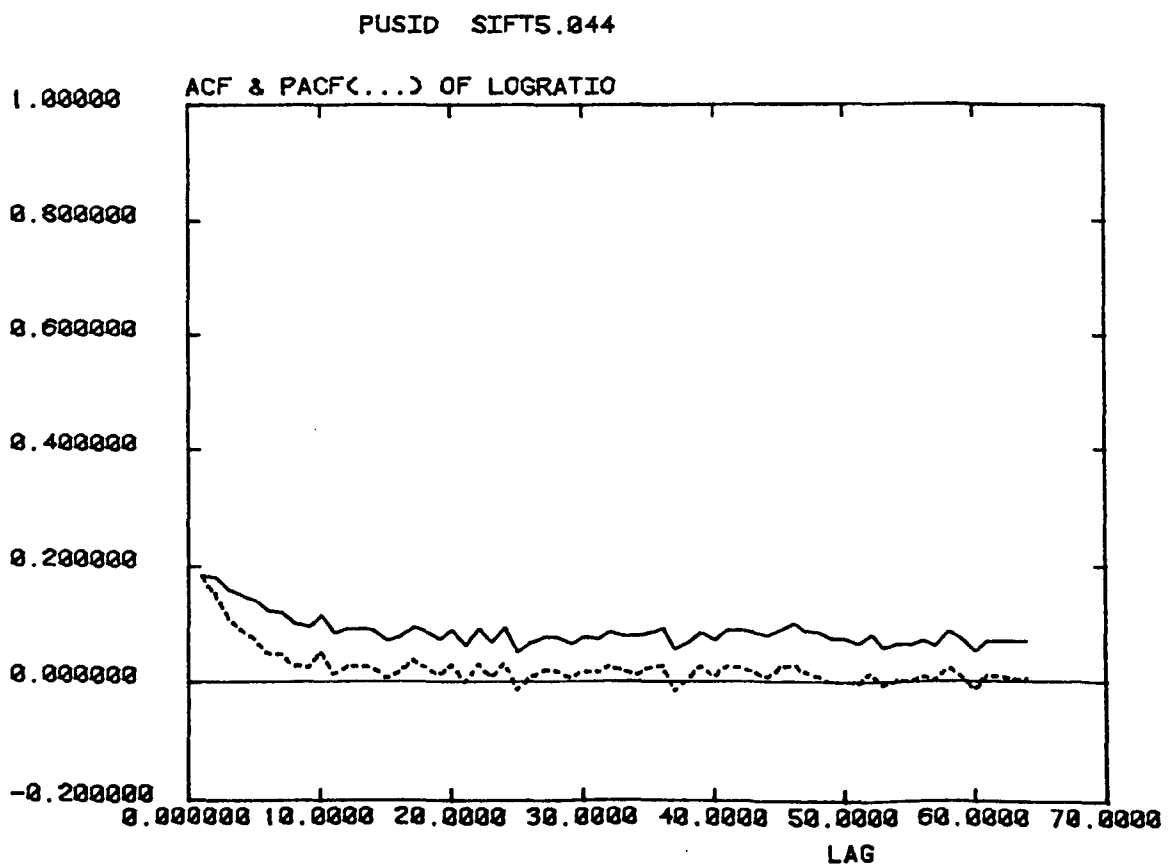


FIGURE 7.10 (g)



acf and pacf of log Ratio Data : SIFT 5·043 (above)
: SIFT 5·044 (below)

FIGURE 7.10 (h)



acf and pacf of Log Ratio Data for SIFT 5·046

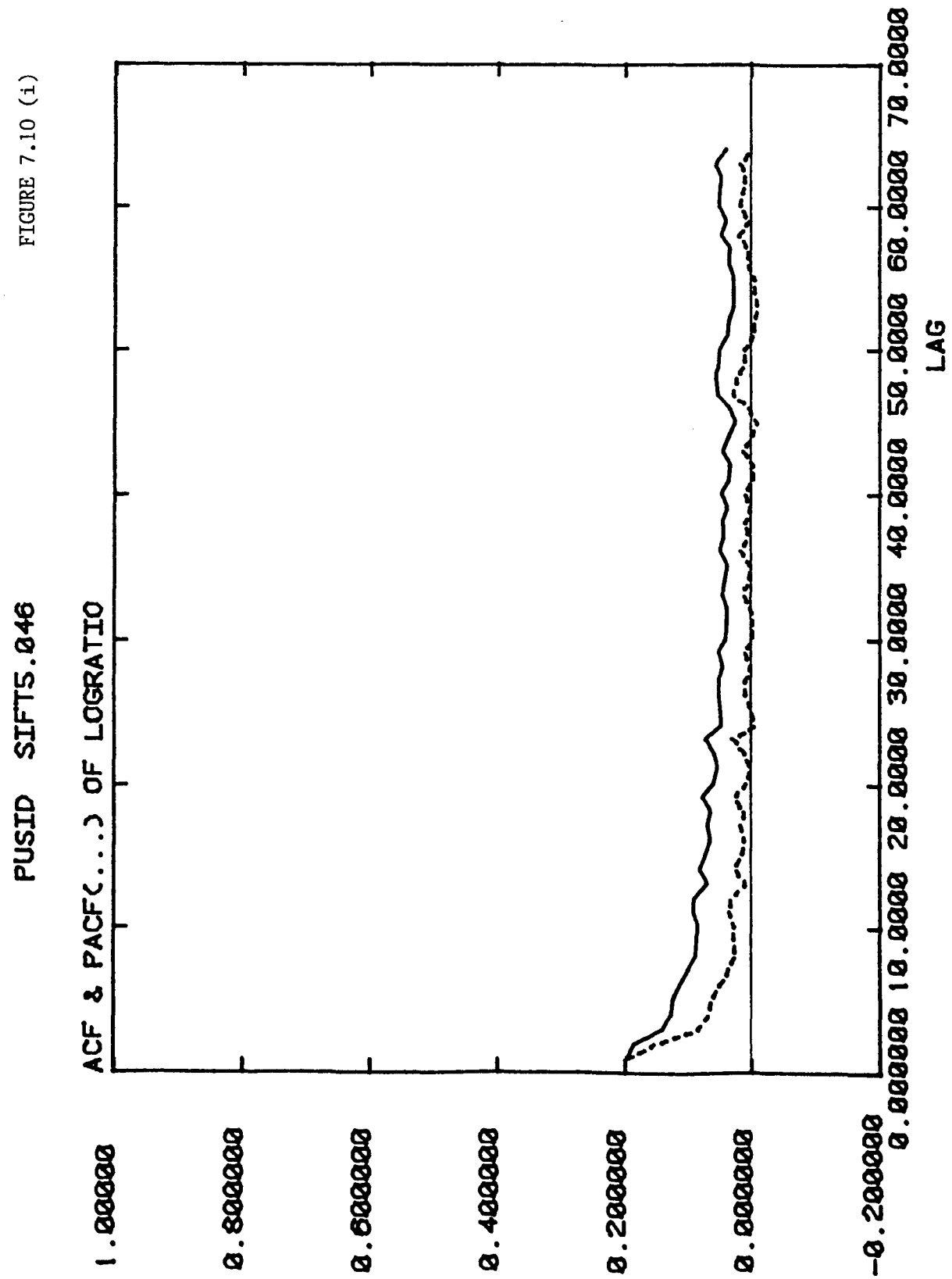
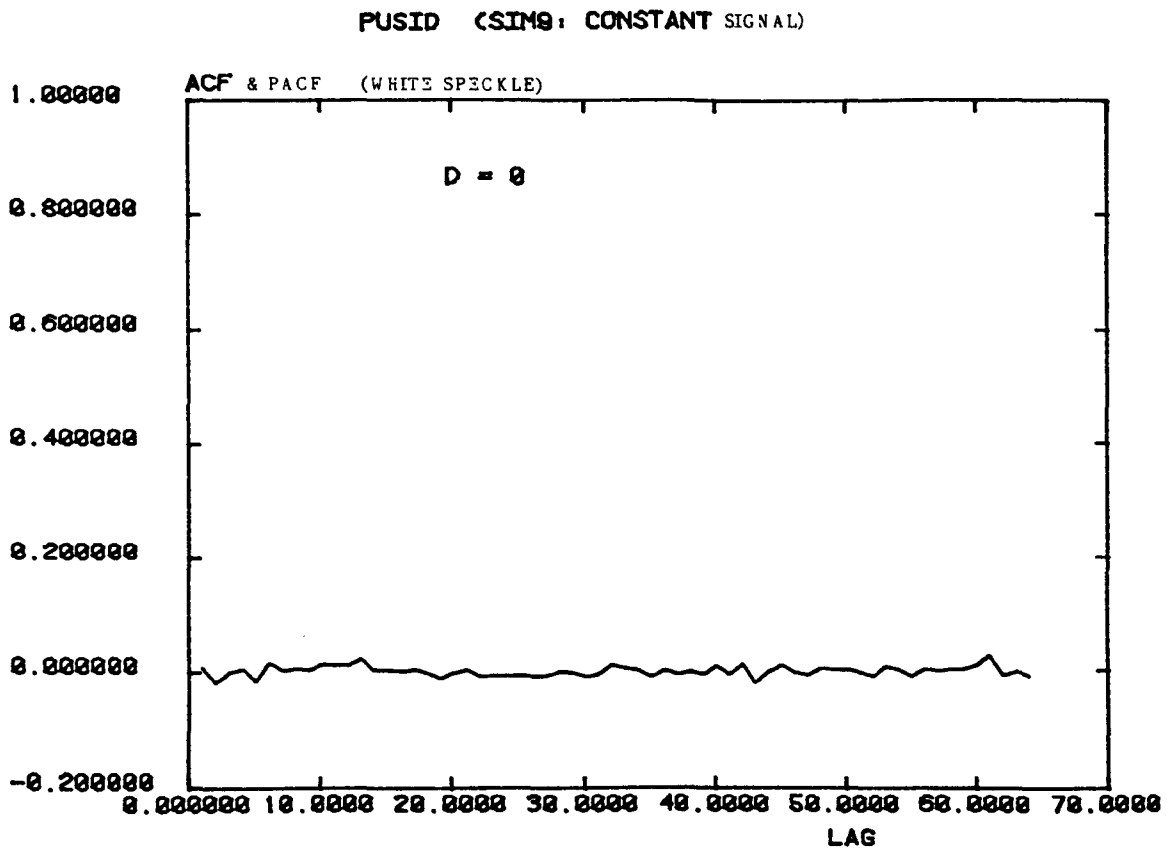


FIGURE 7.10 (i)

FIGURE 7.11 (a)



acf and pacf of Simulated Lidar Data: Uncorrelated Speckle (above)
: Correlated Speckle (below)

FIGURE 7.11 (b)

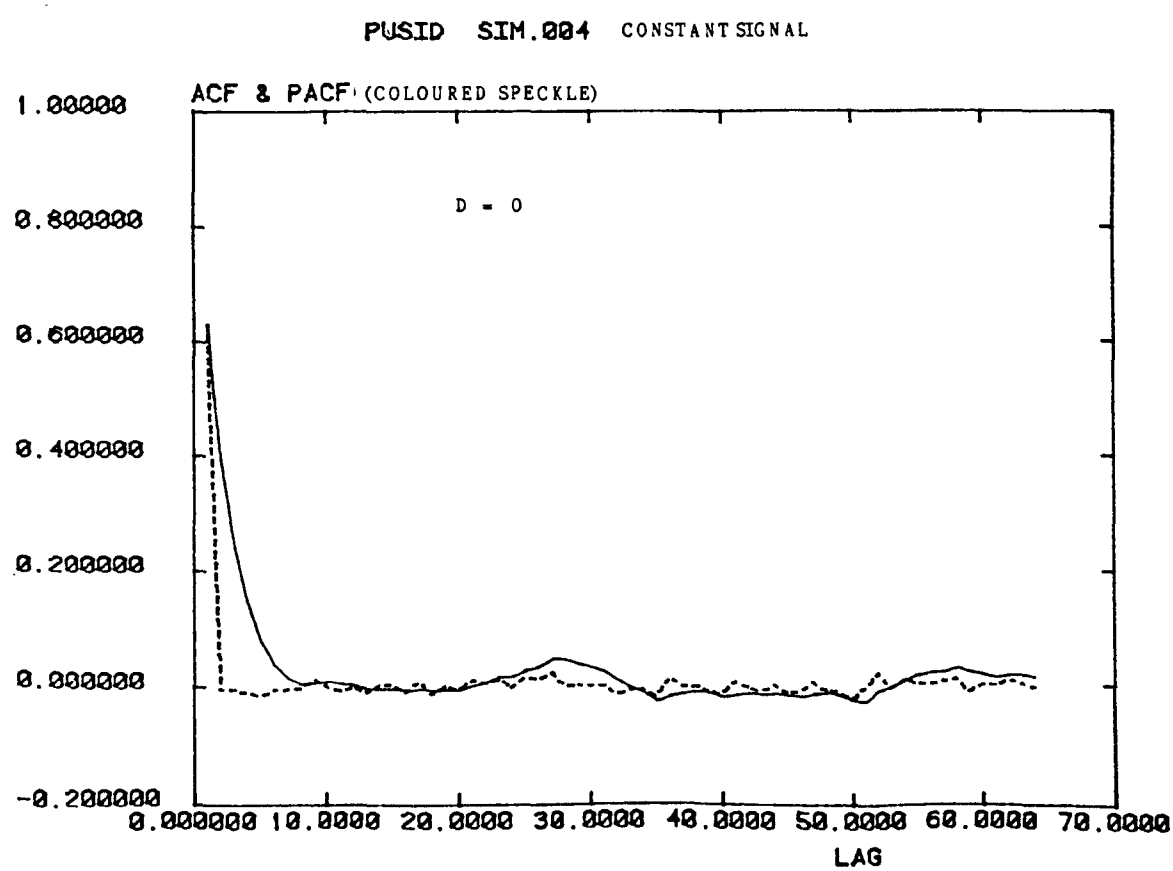
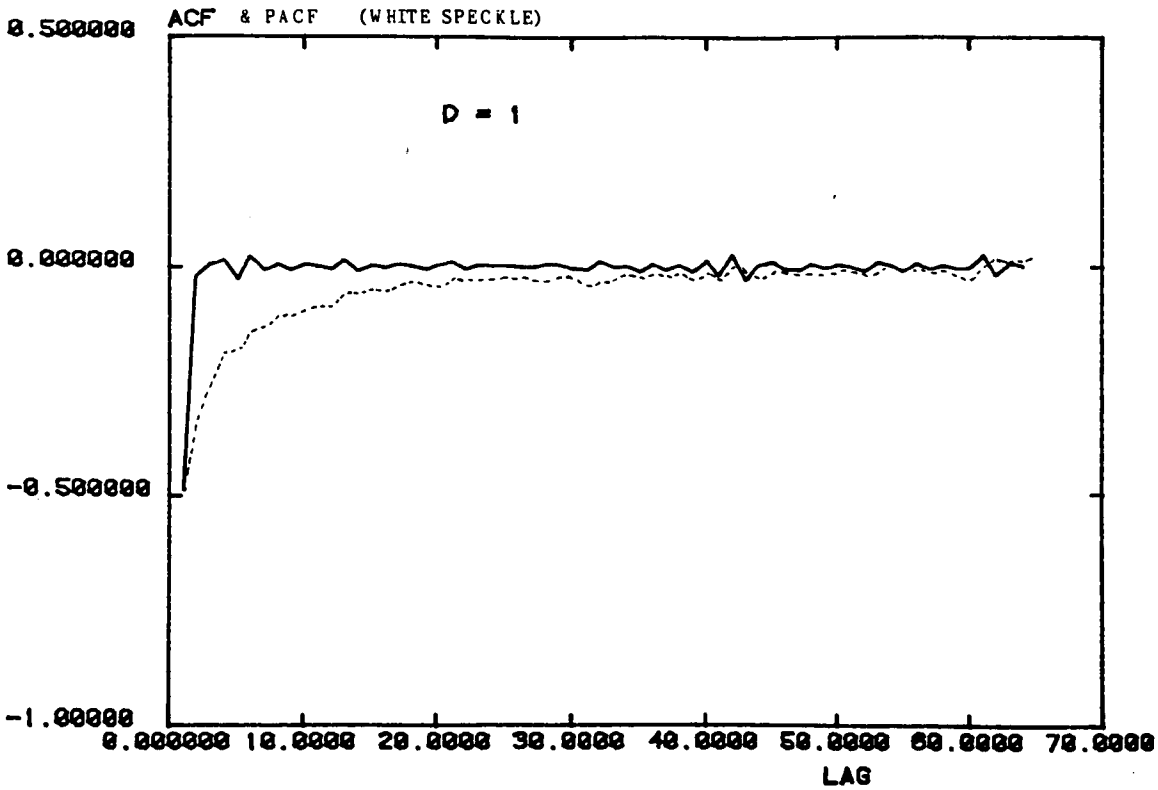


FIGURE 7.11 (c)

PUSID (SIM9: CONSTANT SIGNAL)



acf and pacf of Simulated Lidar Data with First Order Differencing:
Uncorrelated Speckle (above)
Correlated Speckle (below)

FIGURE 7.11 (d)

PUSID SIM.004 CONSTANT SIGNAL

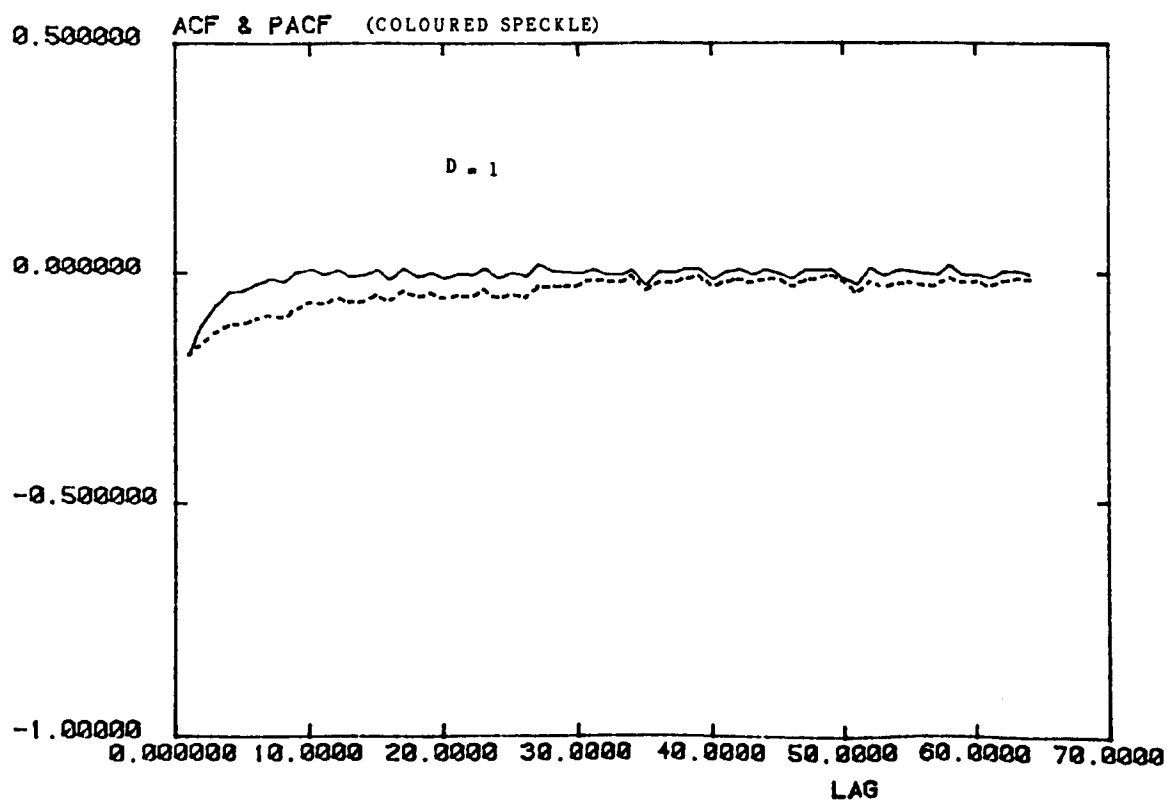
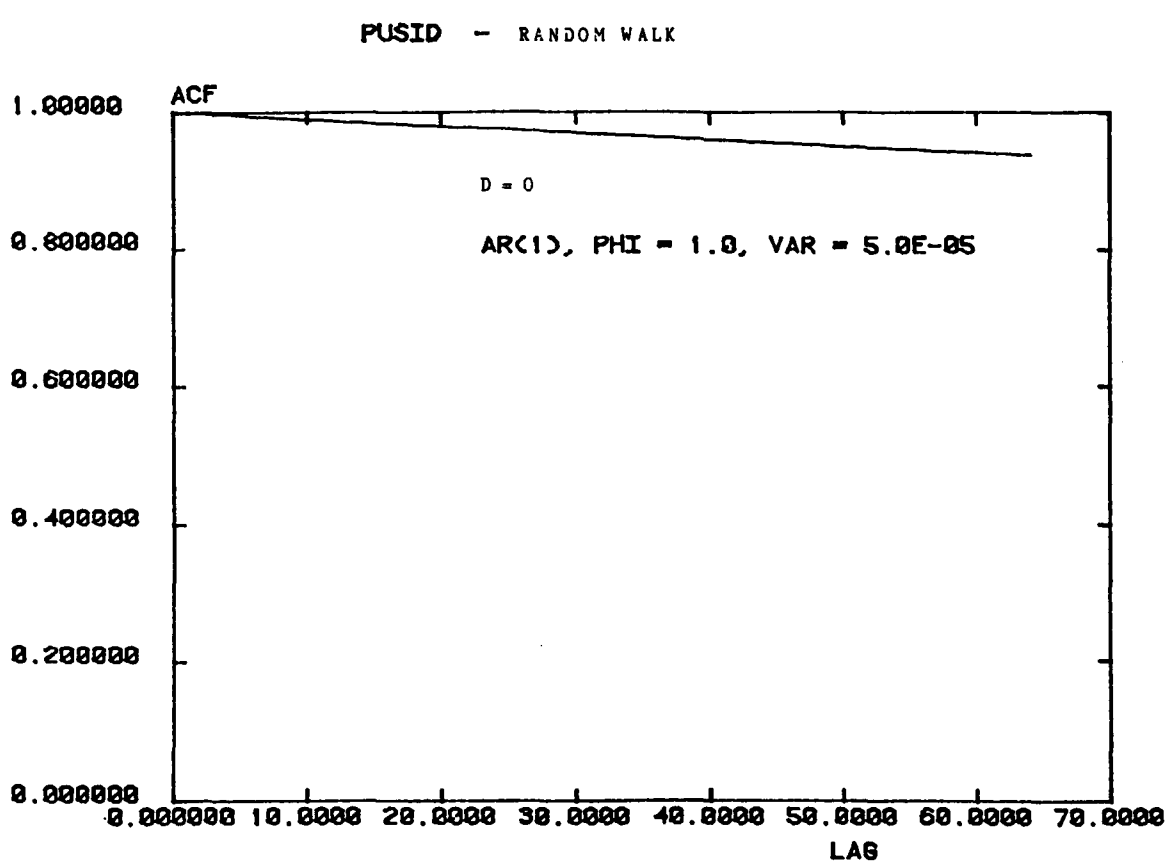


FIGURE 7.11 (e)



acf of a Pure Random Walk Signal (above)
pacf of a Pure Random Walk Signal (below)

FIGURE 7.11 (f)

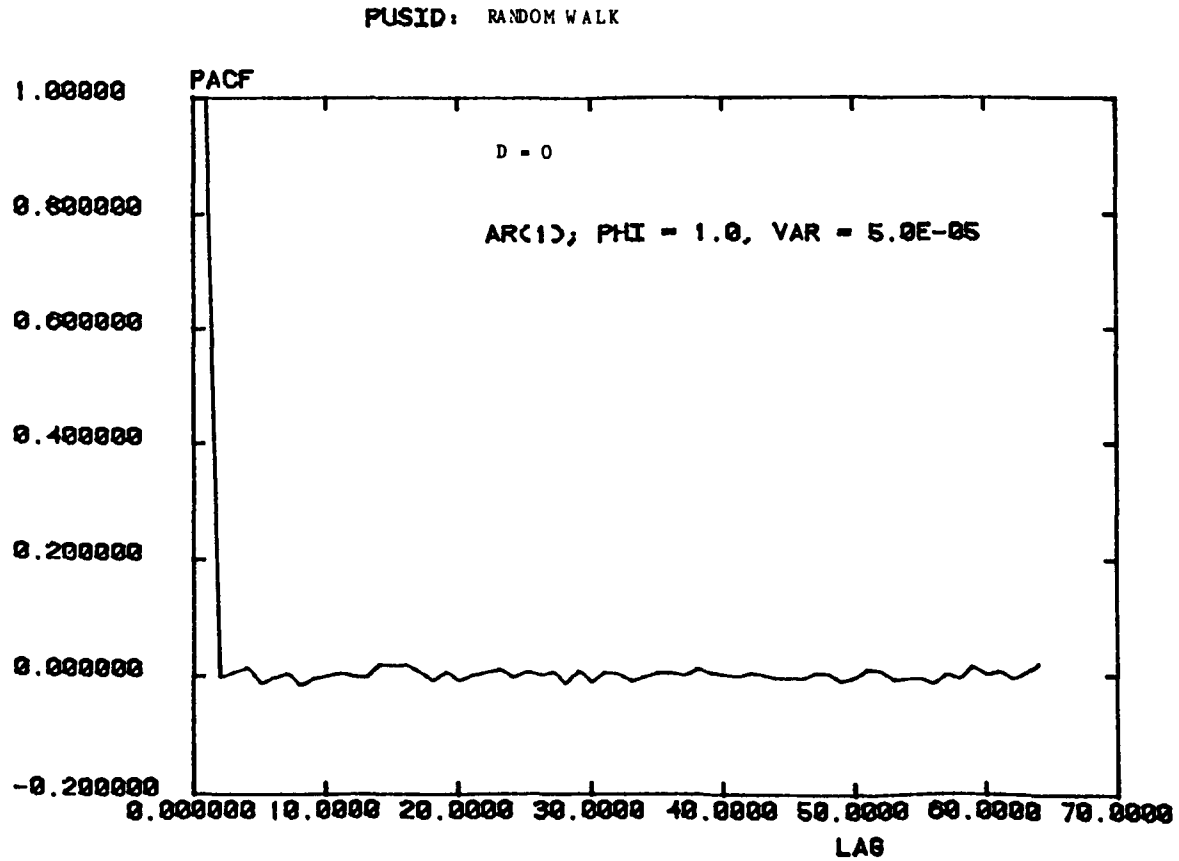
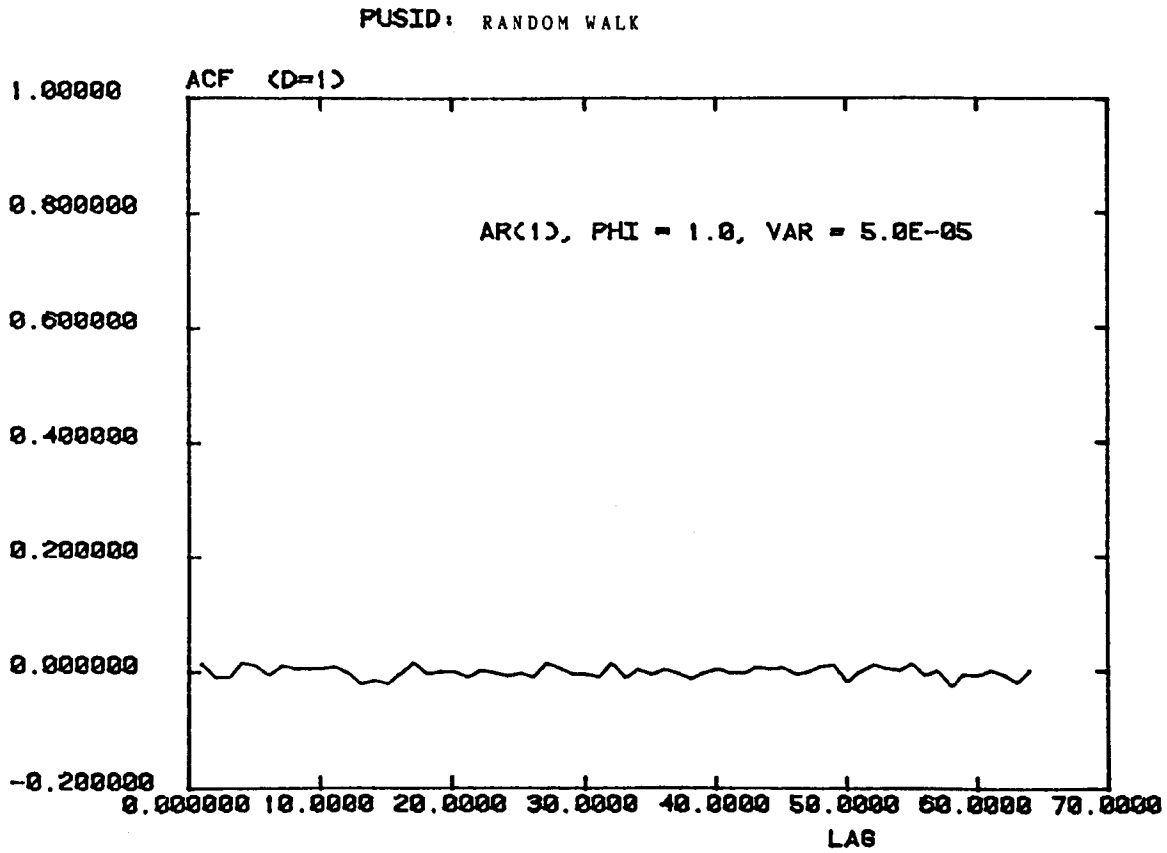


FIGURE 7.11 (g)



act of a Differenced Random Walk (above)
pact of a Differenced Random Walk (below)

FIGURE 7.11 (h)

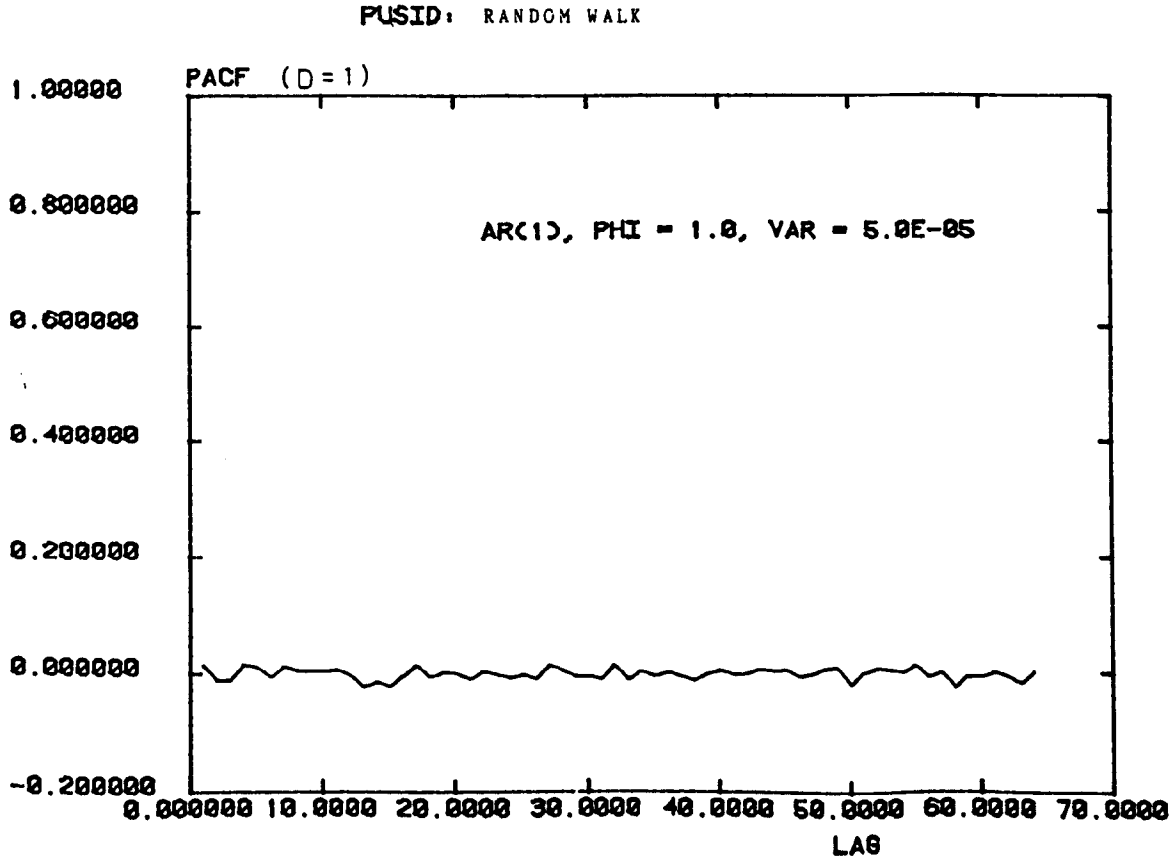
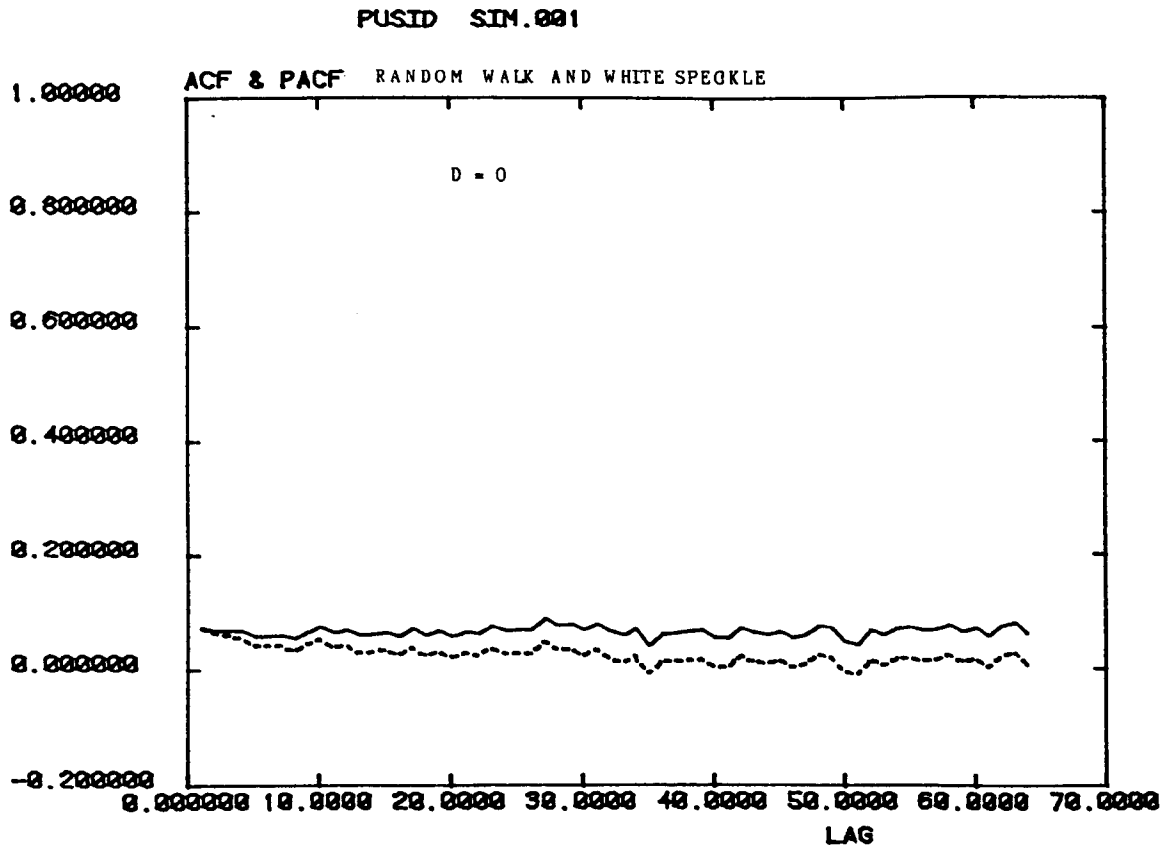


FIGURE 7.11 (i)



acf and pacf of Simulated Data using a Random Walk Signal:
with Uncorrelated Speckle (above)
with Correlated Speckle (below)

FIGURE 7.11 (j)

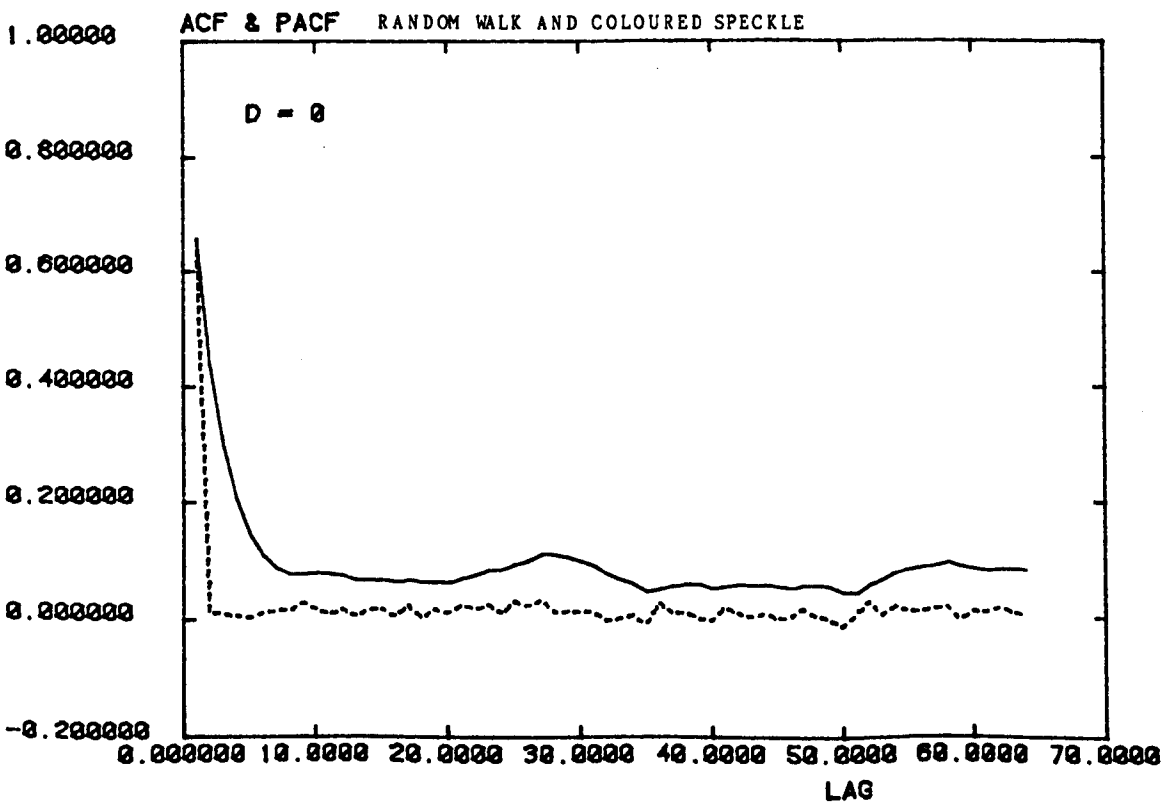
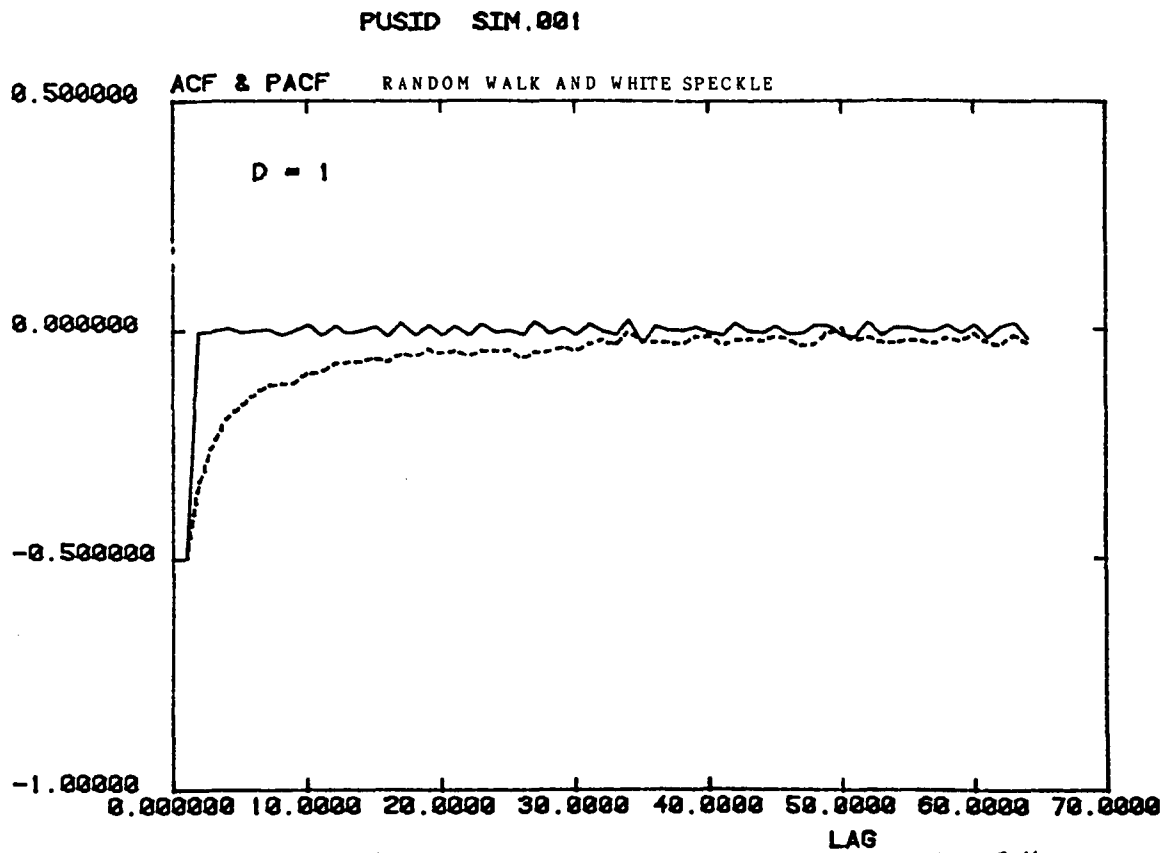


FIGURE 7.11 (k)



acf and pacf of Simulated Data Using a Random Walk Signal with Differencing:
Uncorrelated Speckle (above)
Correlated Speckle (below)

FIGURE 7.11 (l)

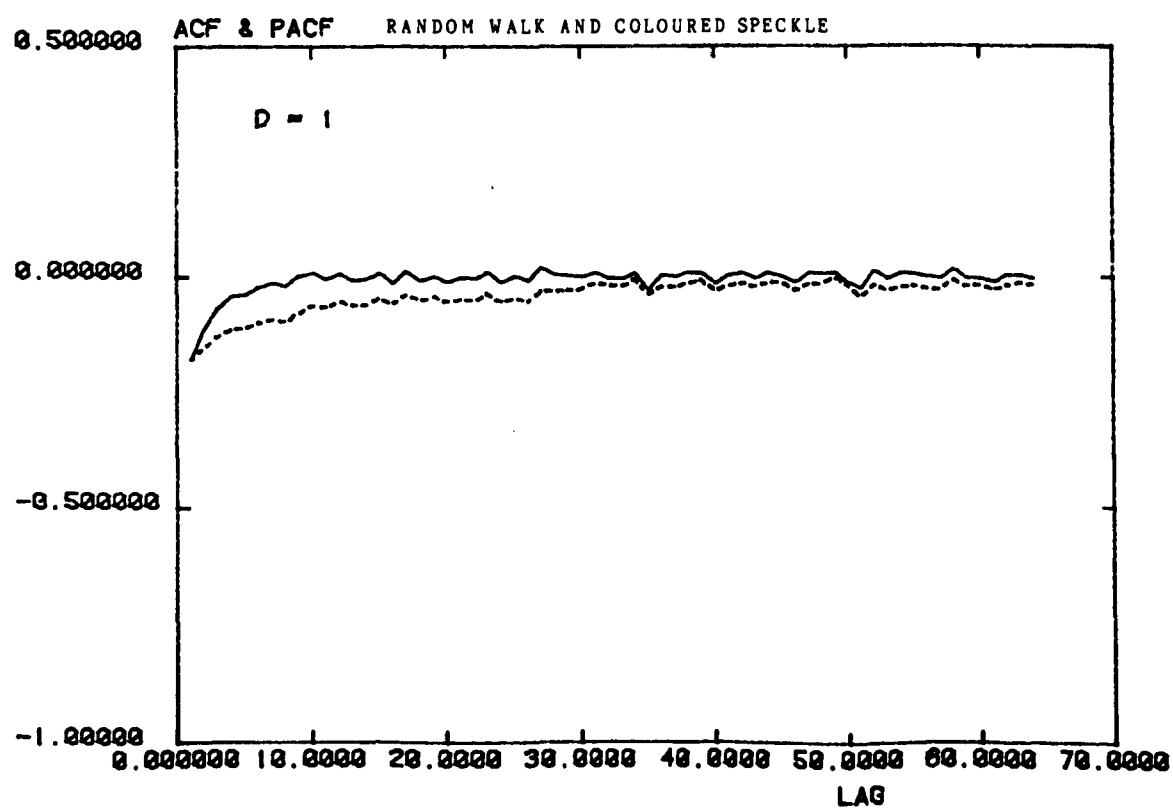
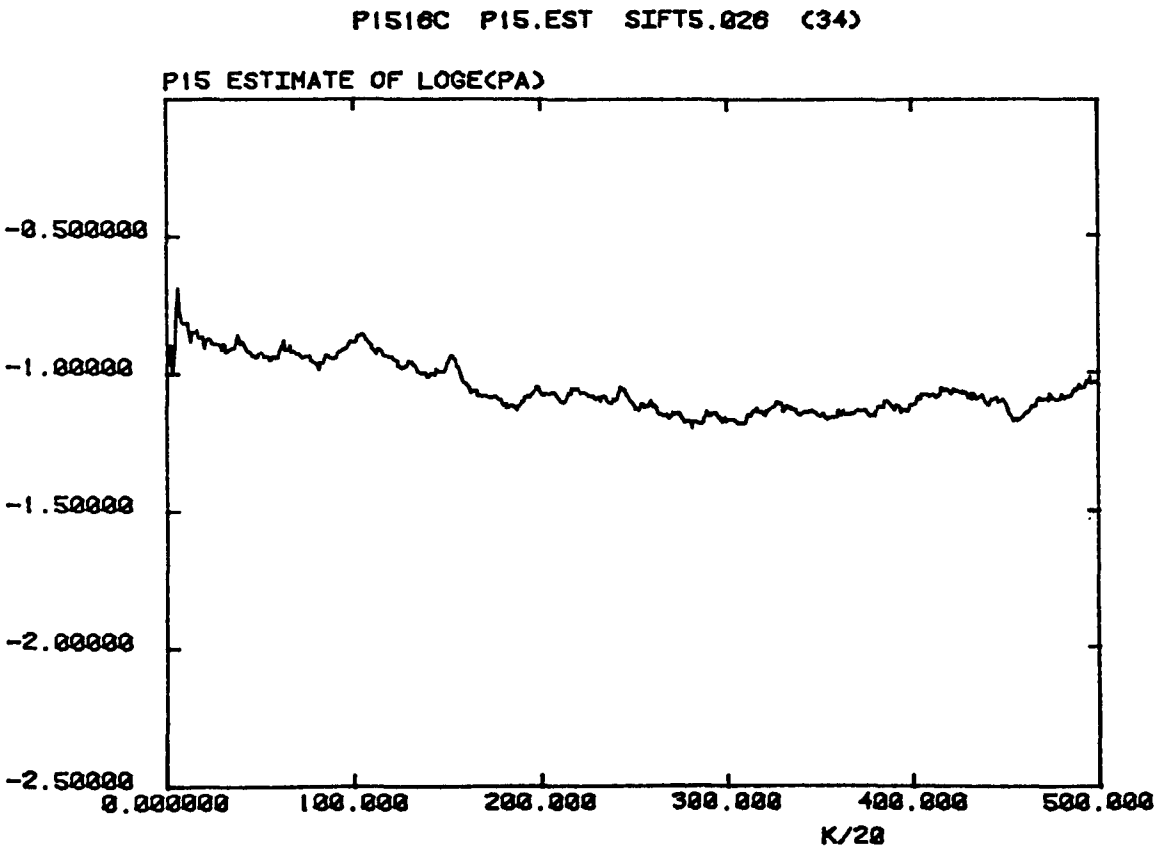


Figure 7.12 (a)



P15 Estimates of log (Normalised Power) for SIFT 5·026:
Channel A (above)
Channel B (below)

Figure 7.12 (b)

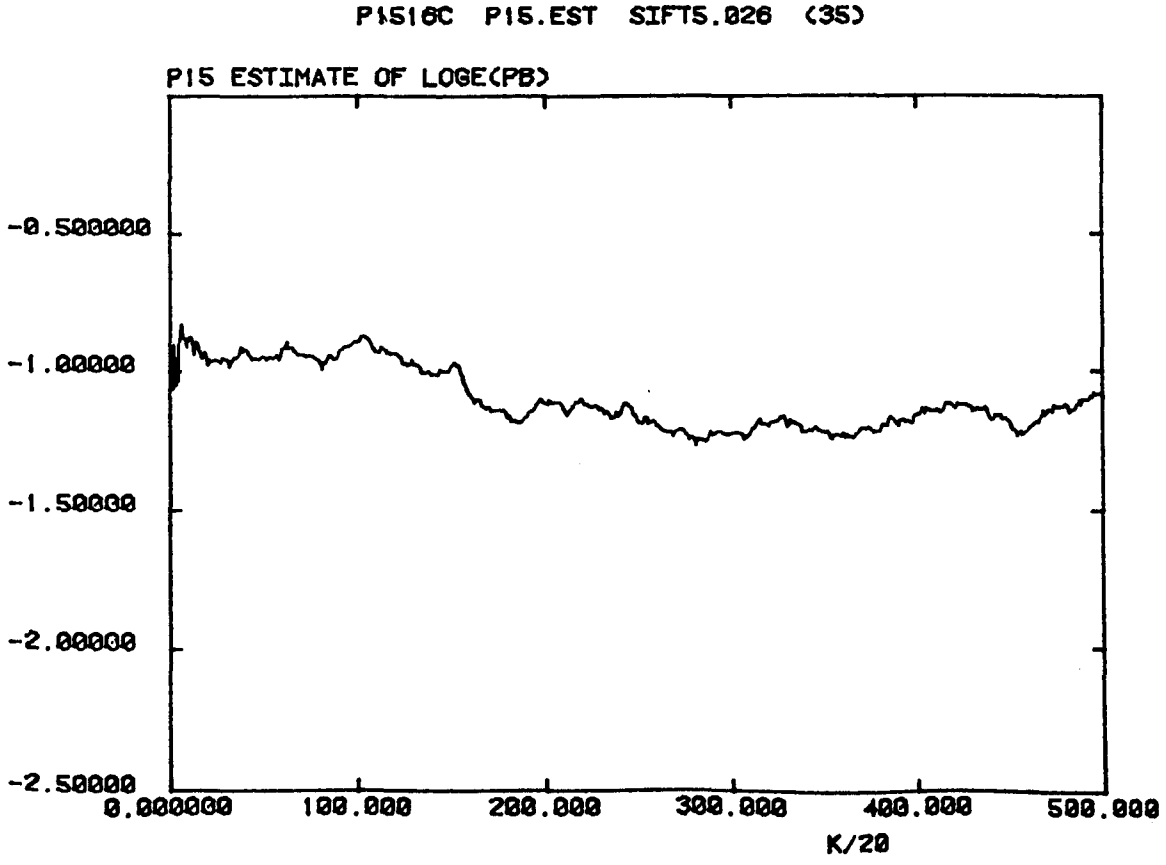
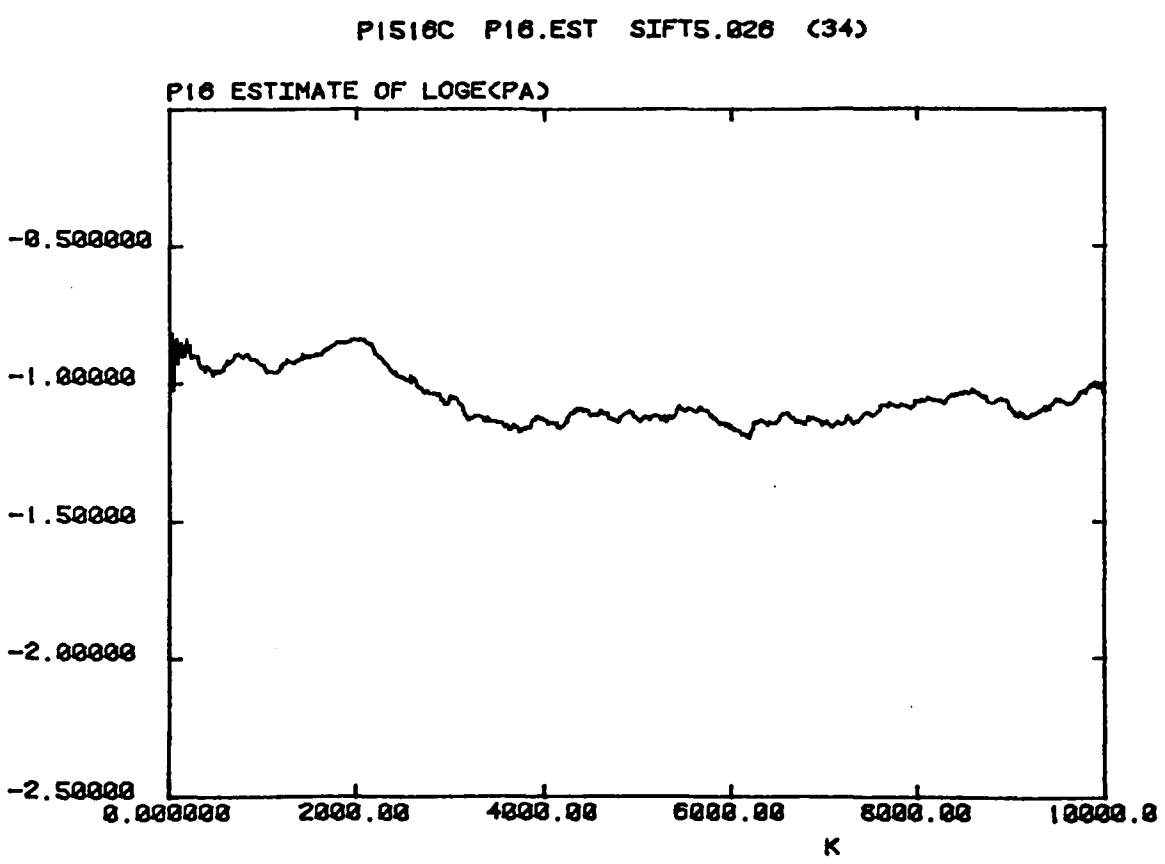


Figure 7.12 (c)



P16 Estimate of Log (Normalised Power) for SIFT 5·026:
Channel A (above)
Channel B (below)

Figure 7.12 (d)

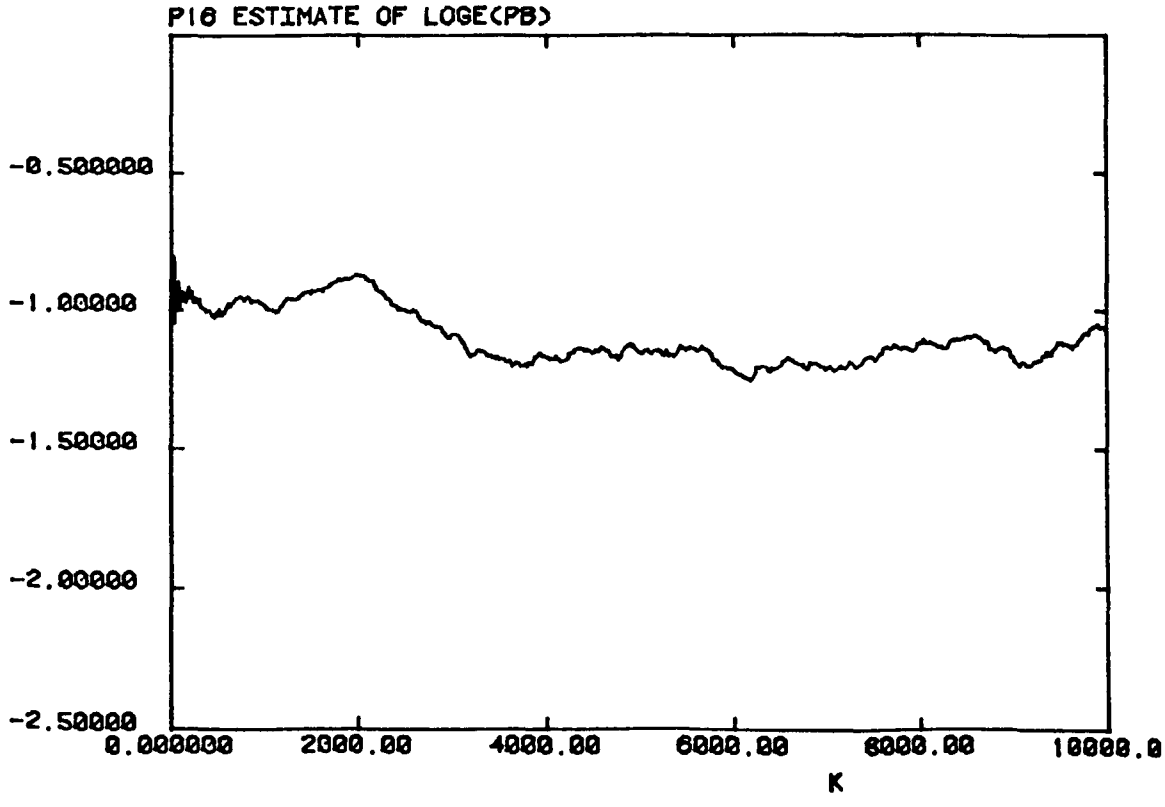
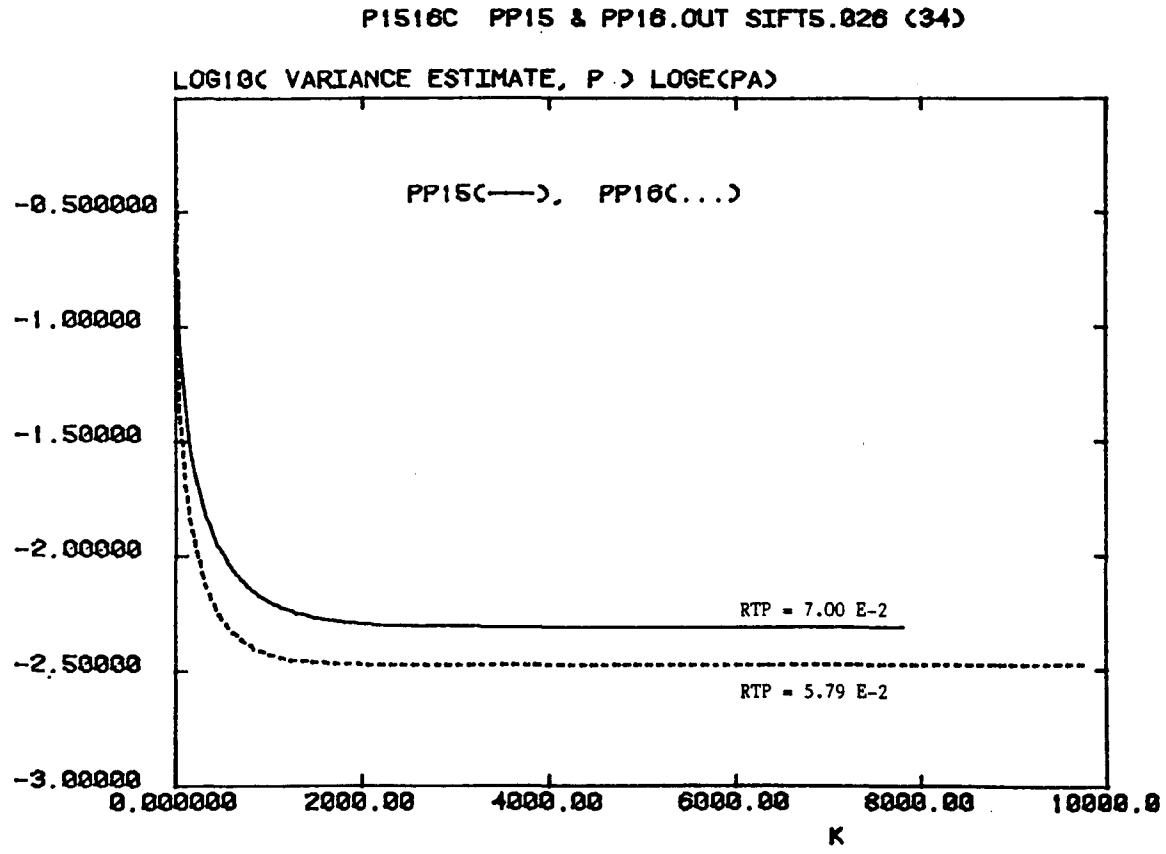


Figure 7.12 (e)



Error Covariance Estimates for SIFT 5·026:
Channel A (above)
Channel B (below)

Figure 7.12 (f)

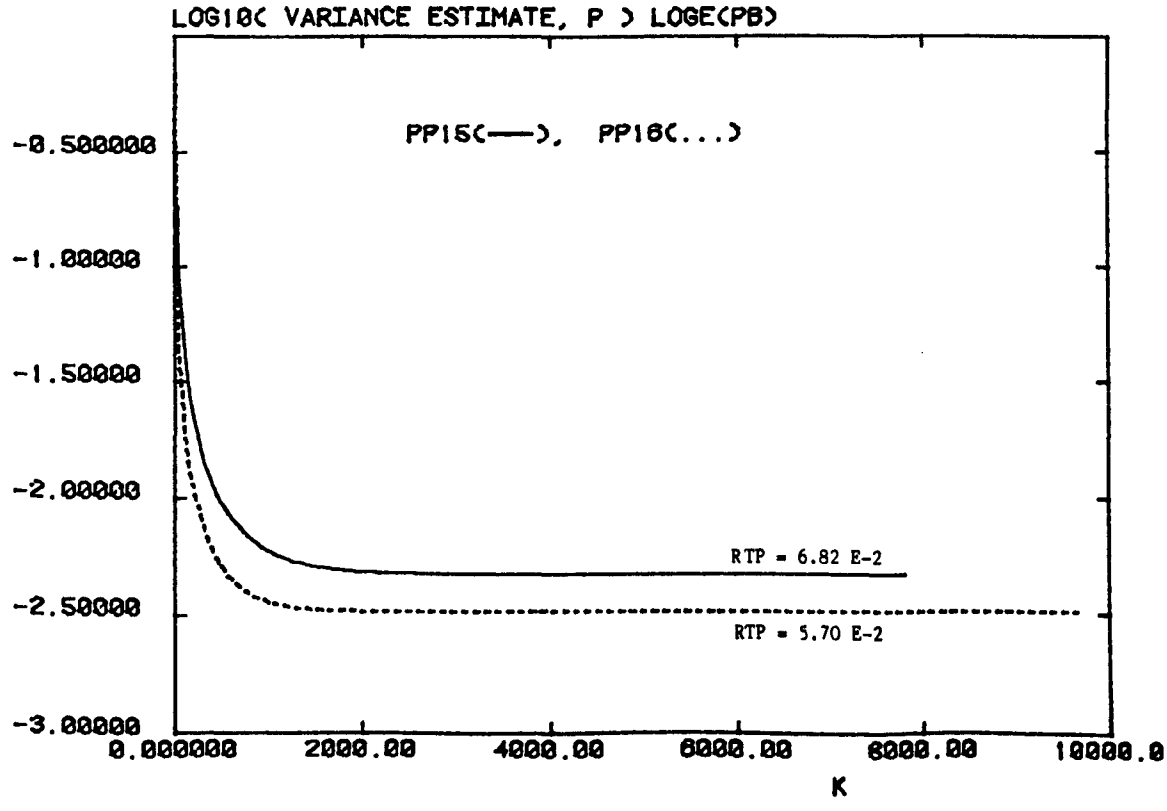
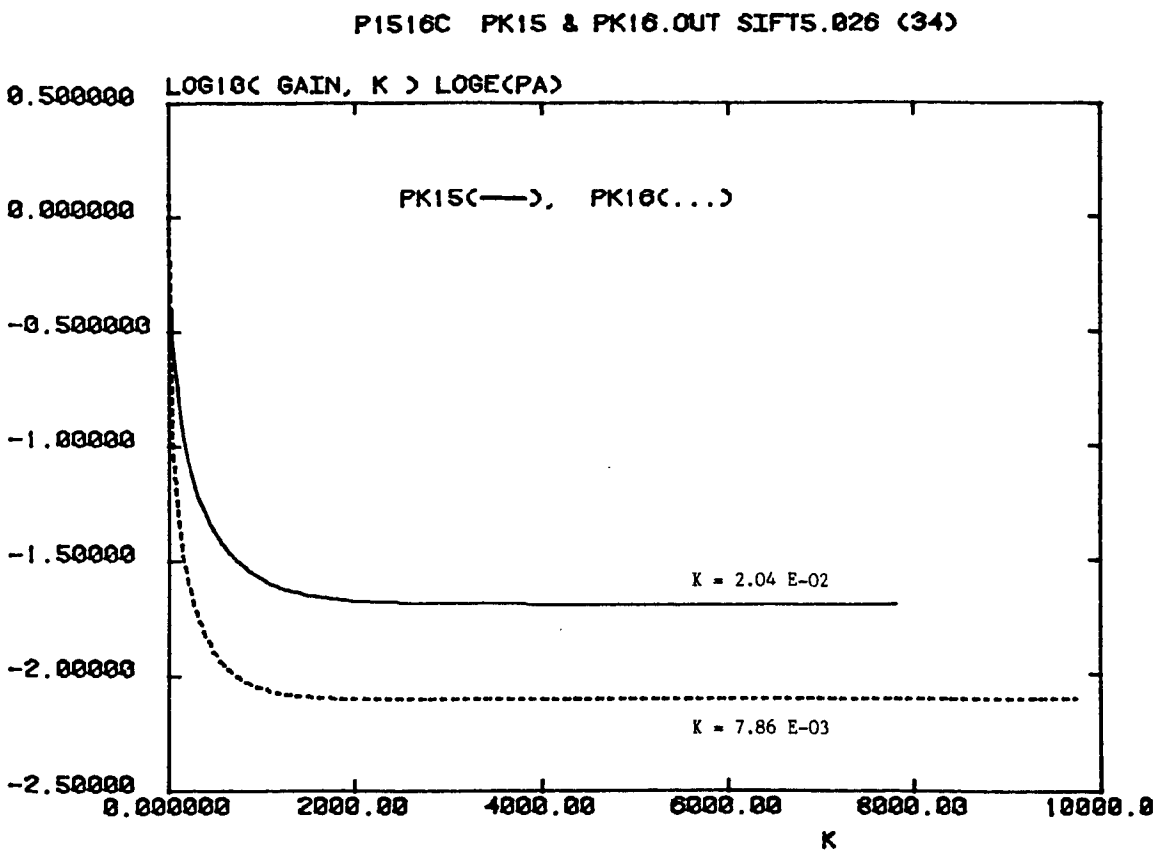


Figure 7.12 (g)



Kalman Gains used with SIFT 5·026: Channel A (above)
: Channel B (below)

Figure 7.12 (h)

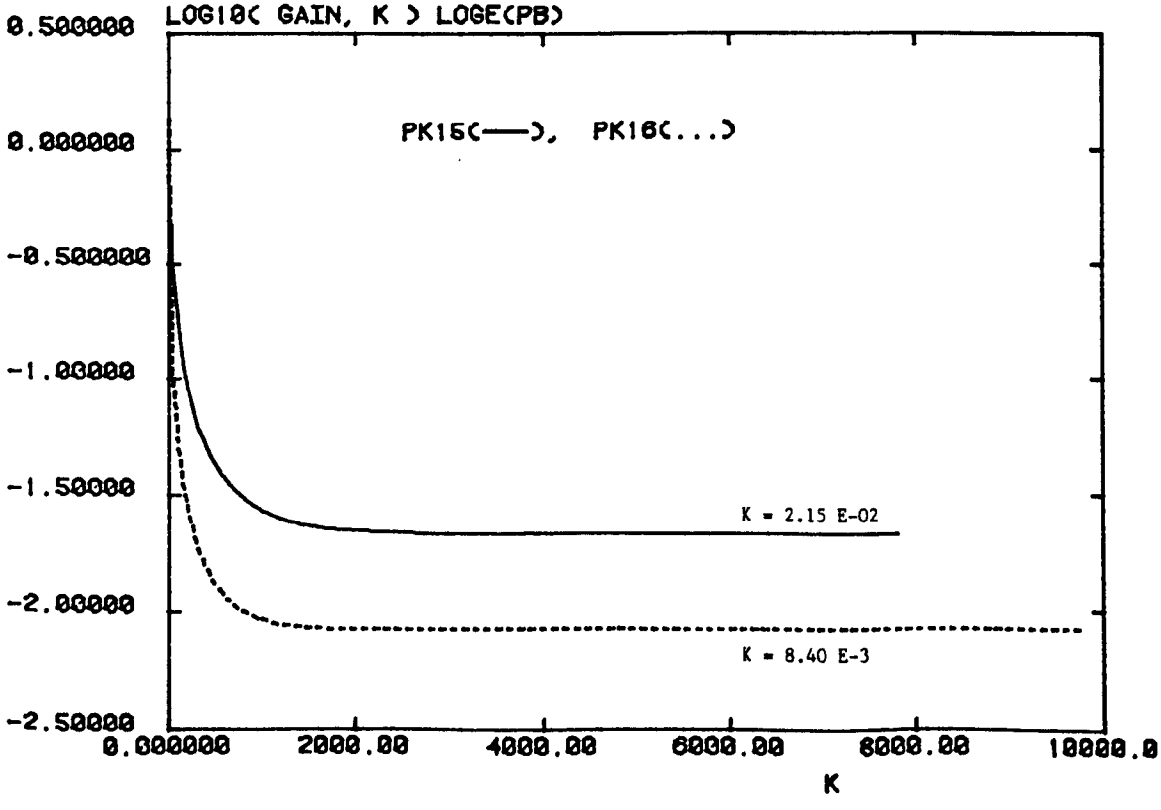
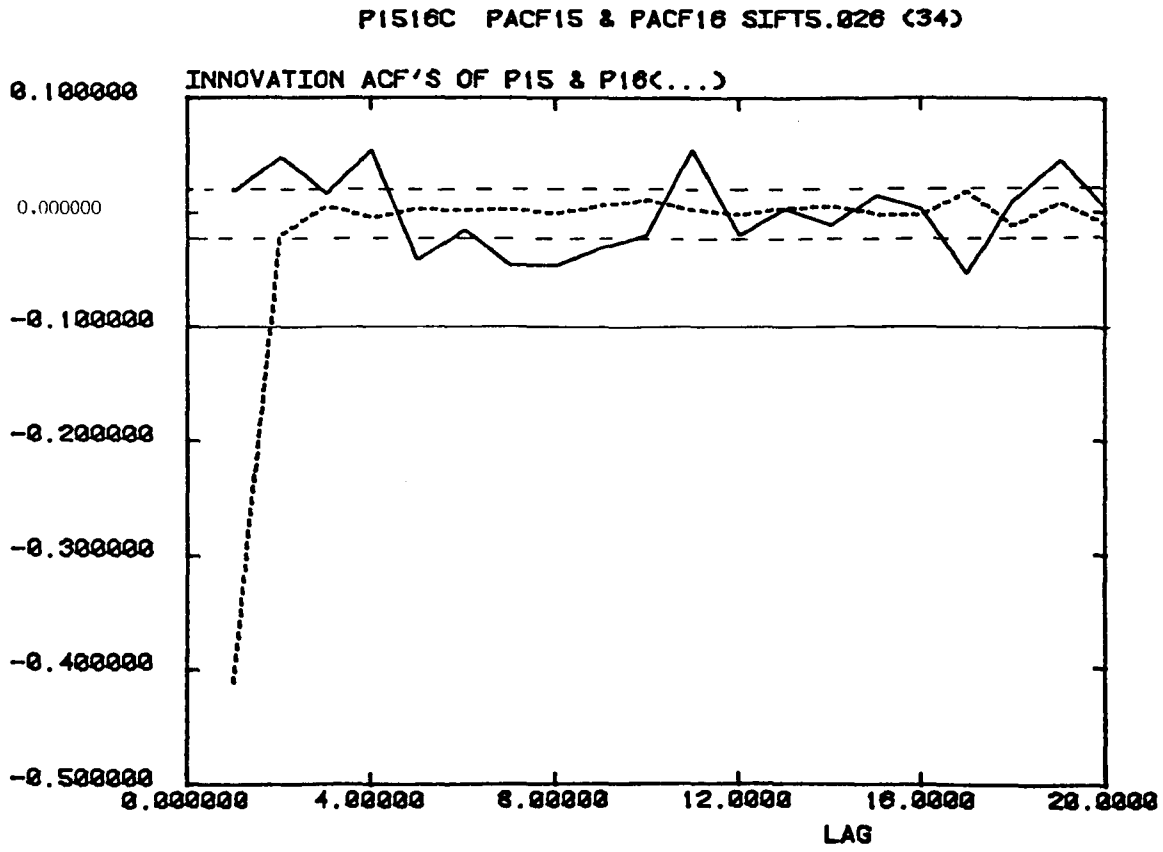


Figure 7.12 (i)



Innovation Sequence acfs for SIFT 5·026 : Channel A (above)
: Channel B (below)

Figure 7.12 (j)

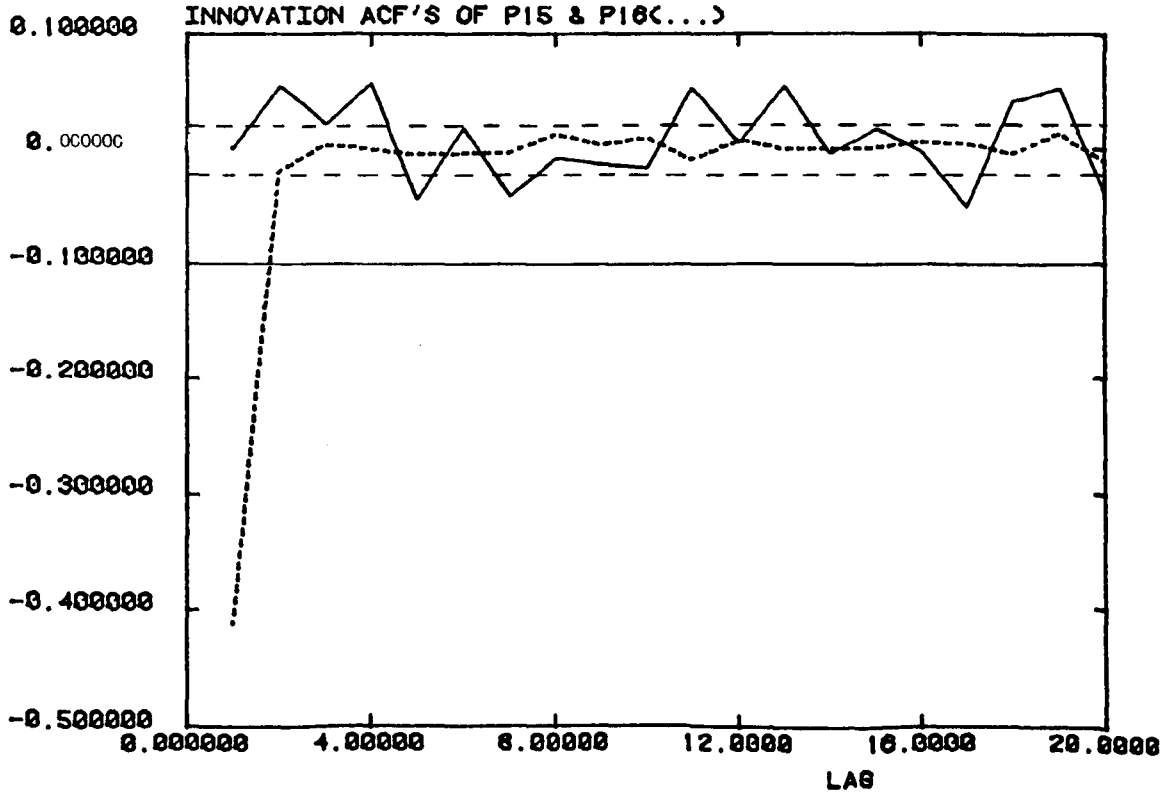
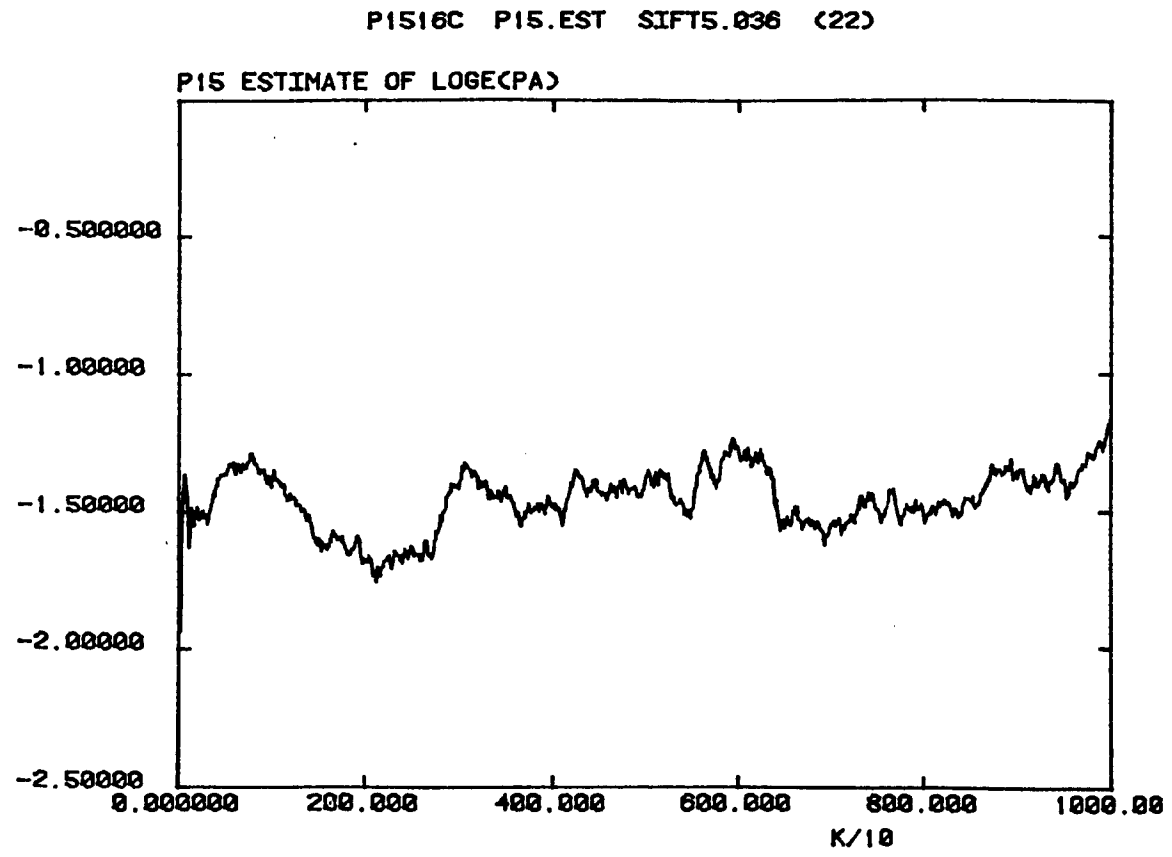


Figure 7.13 (a)



P15 Estimates of log (Normalised Power) for SIFT 5·036:
Channel A (above)
Channel B (below)

Figure 7.13 (b)

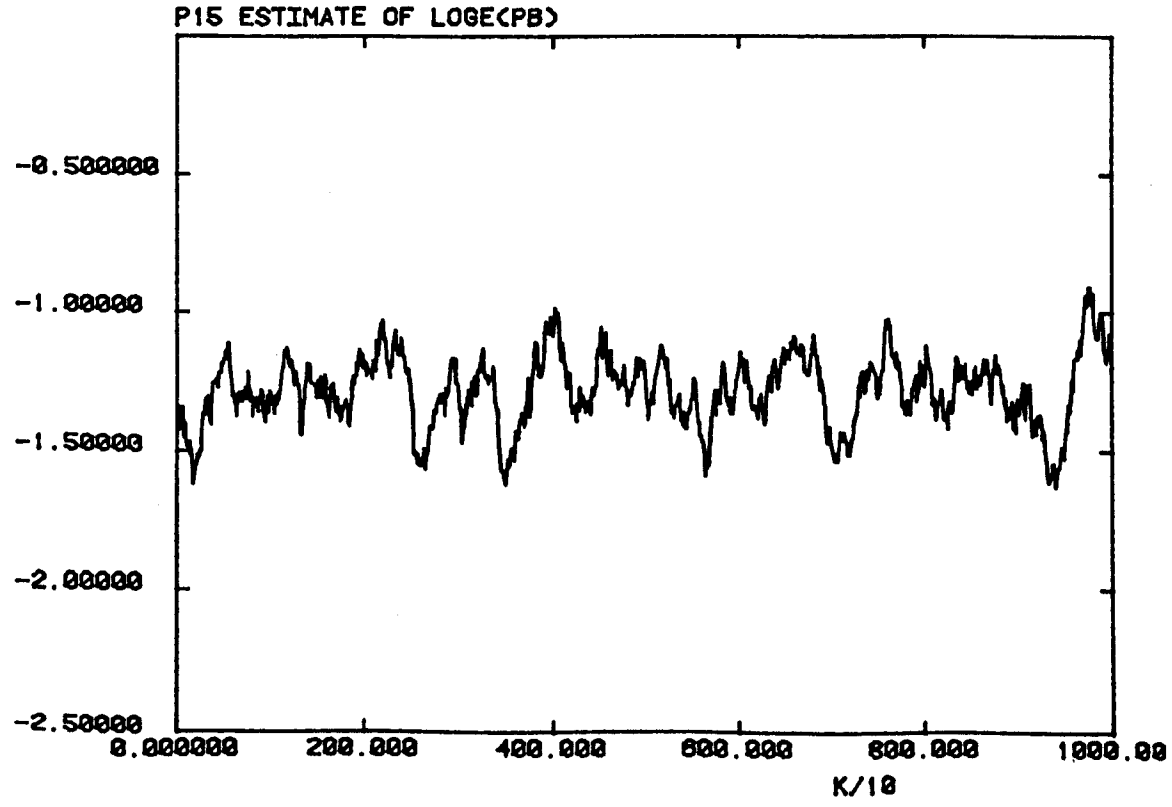
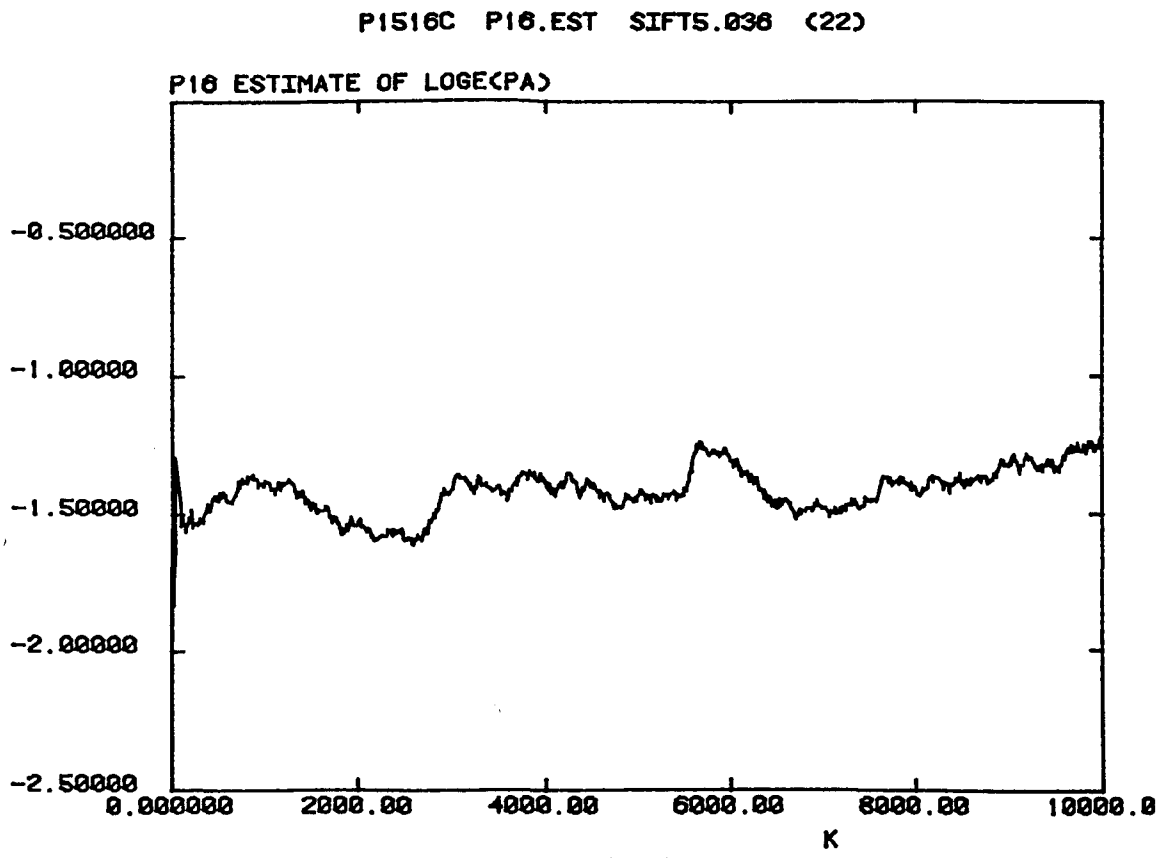


Figure 7.13 (c)



P16 Estimate of Log (Normalised Power) for SIFT 5·036:
Channel A (above)
Channel B (below)

Figure 7.13 (d)

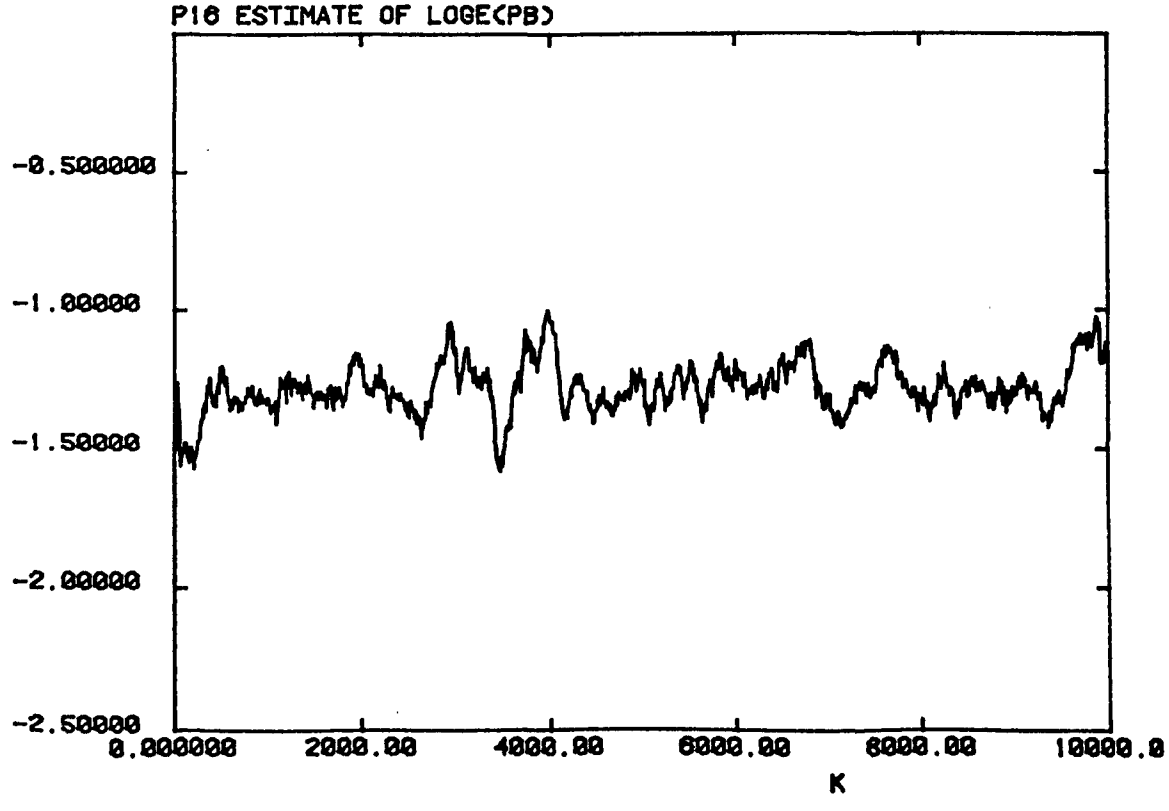
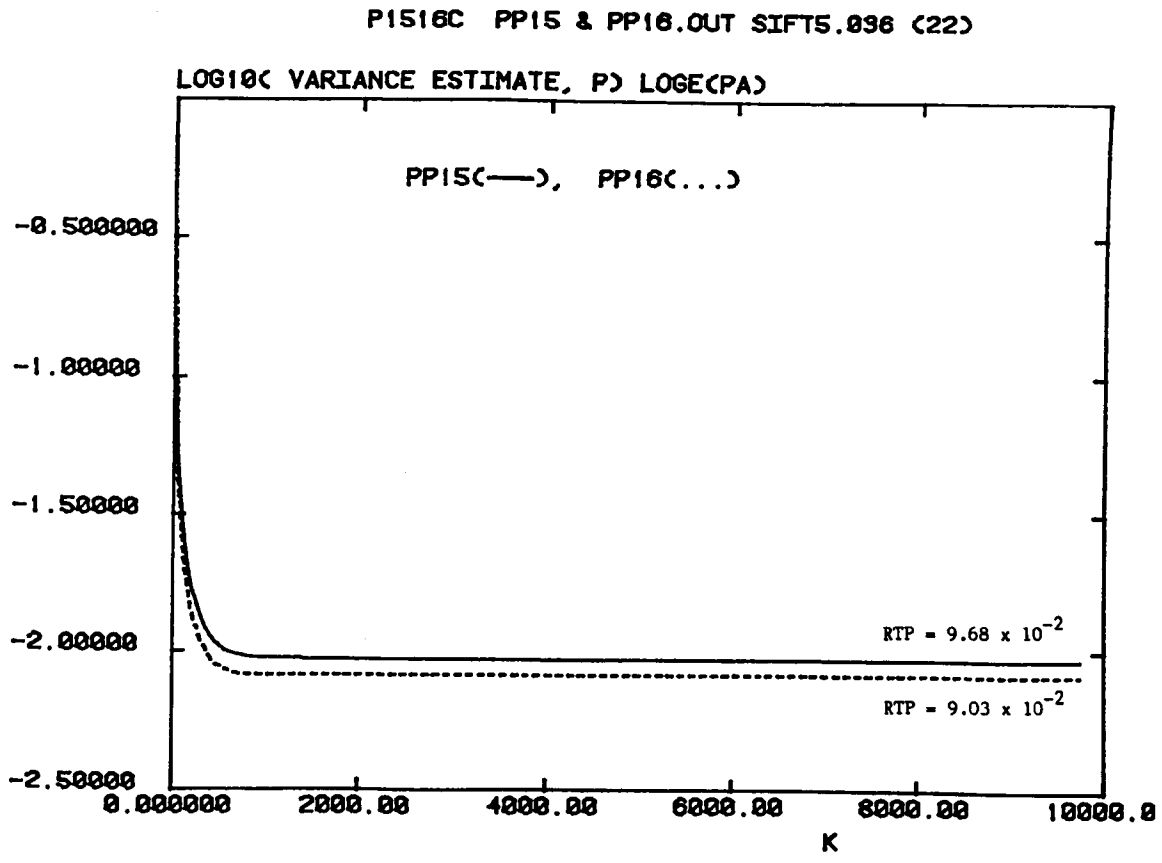


Figure 7.13 (e)



Error Covariance Estimates for SIFT 5·036 : Channel A (above)
: Channel B (below)

Figure 7.13 (f)

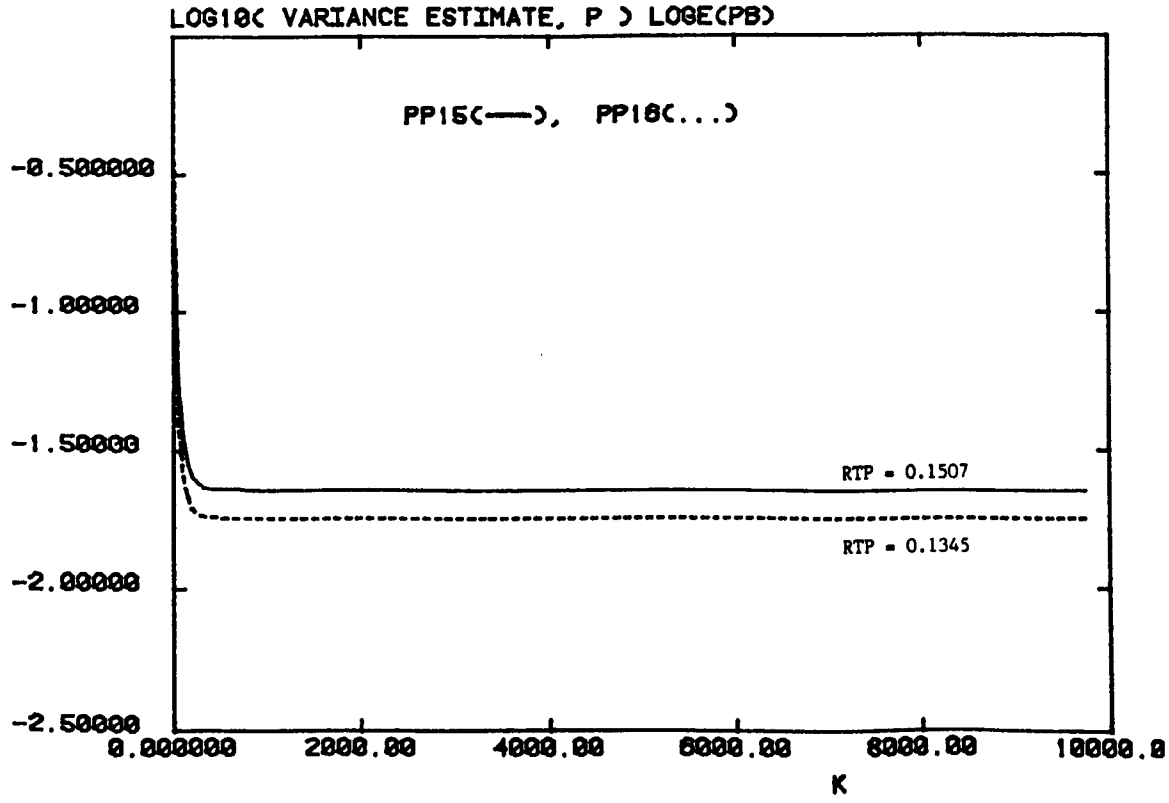
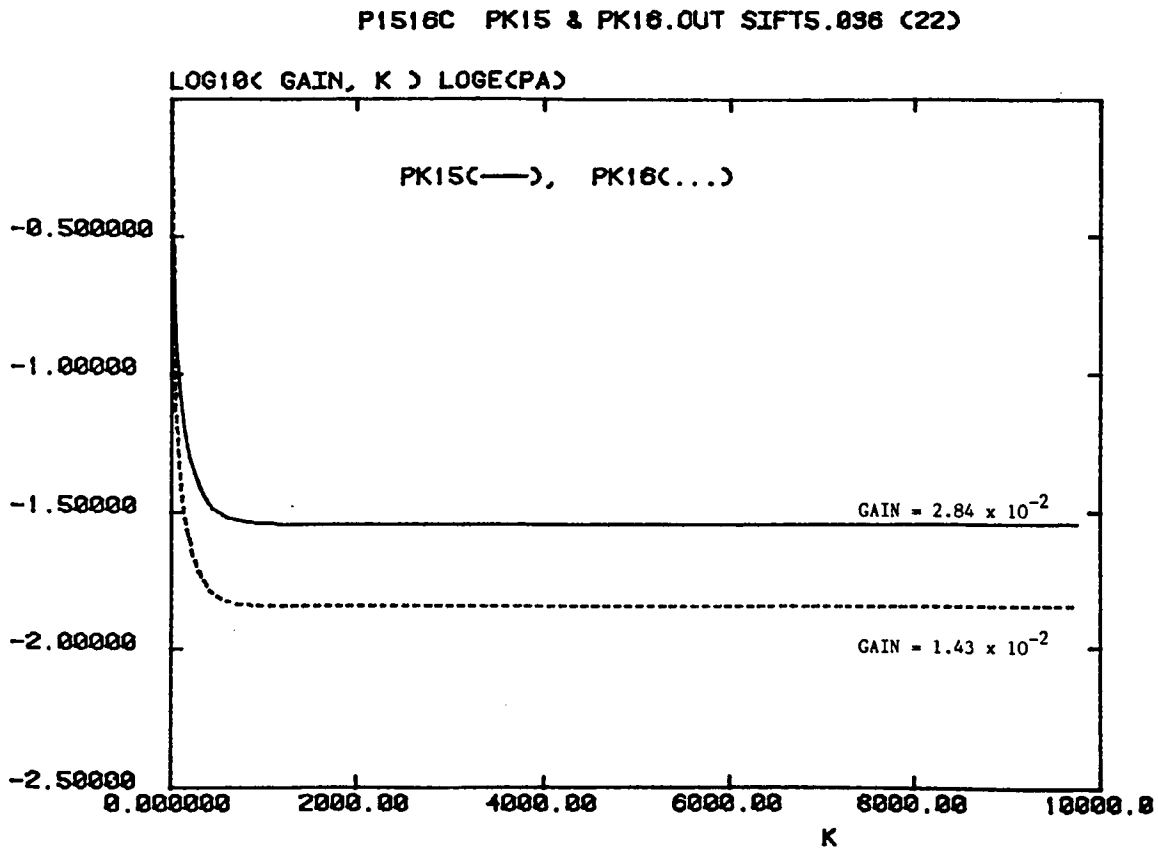


Figure 7.13 (g)



Kalman Gains used with SIFT 5·036: Channel A (above)
: Channel B (below)

Figure 7.13 (h)

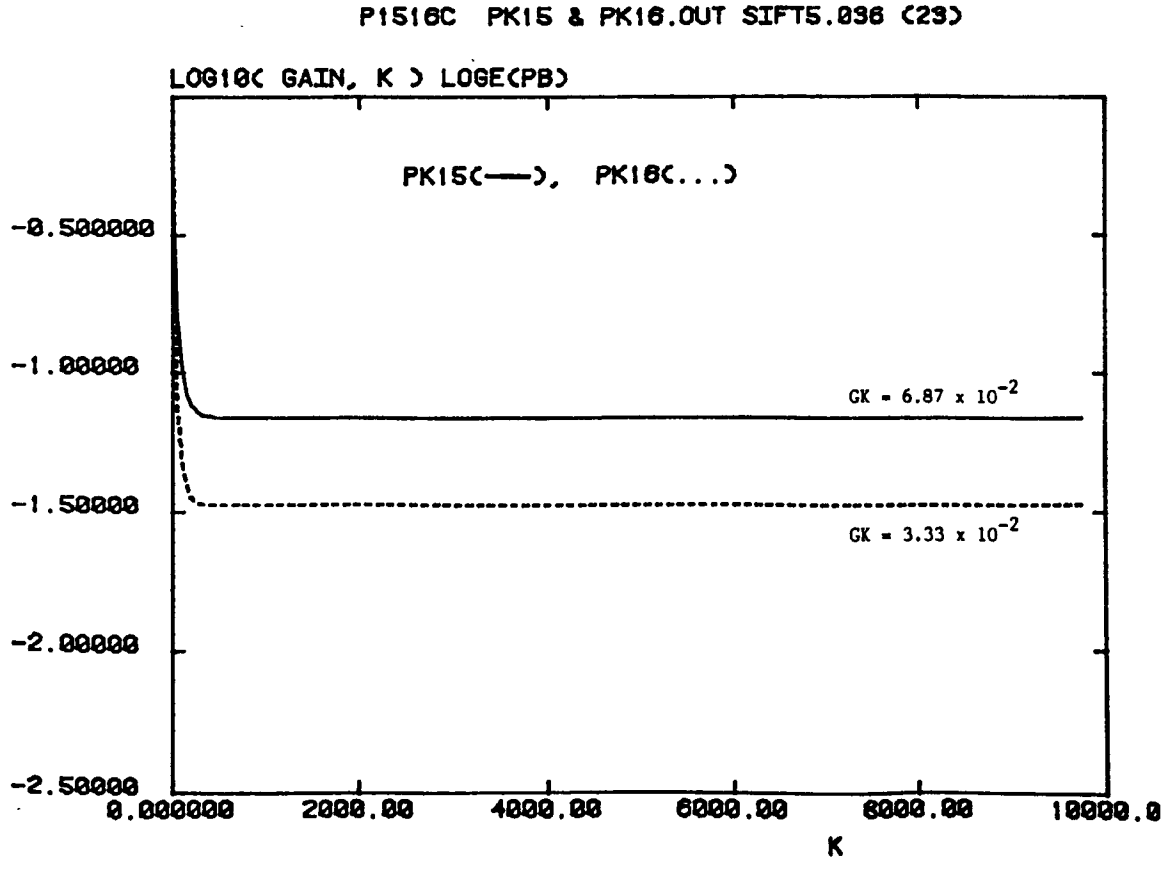
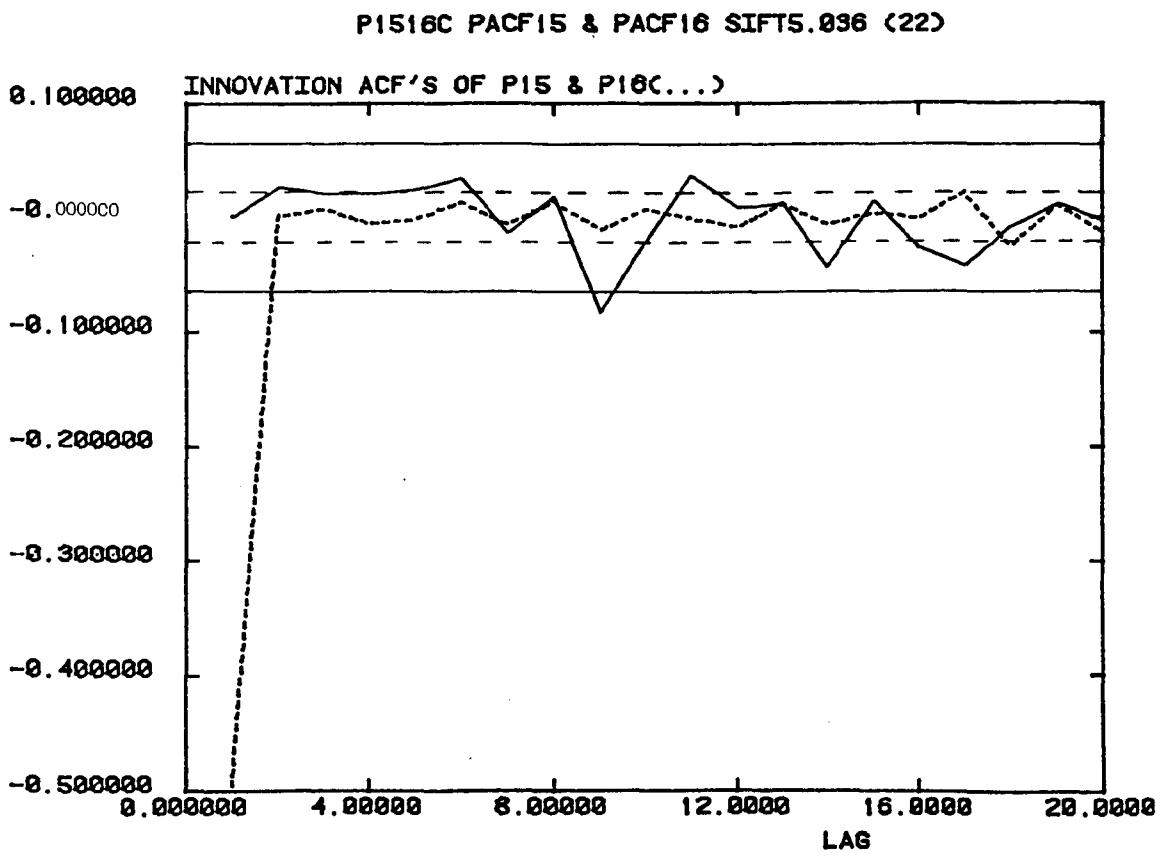


Figure 7.13 (i)



Innovation Sequence acfs for SIFT5·036 : Channel A (above)
: Channel B (below)

Figure 7.13 (j)

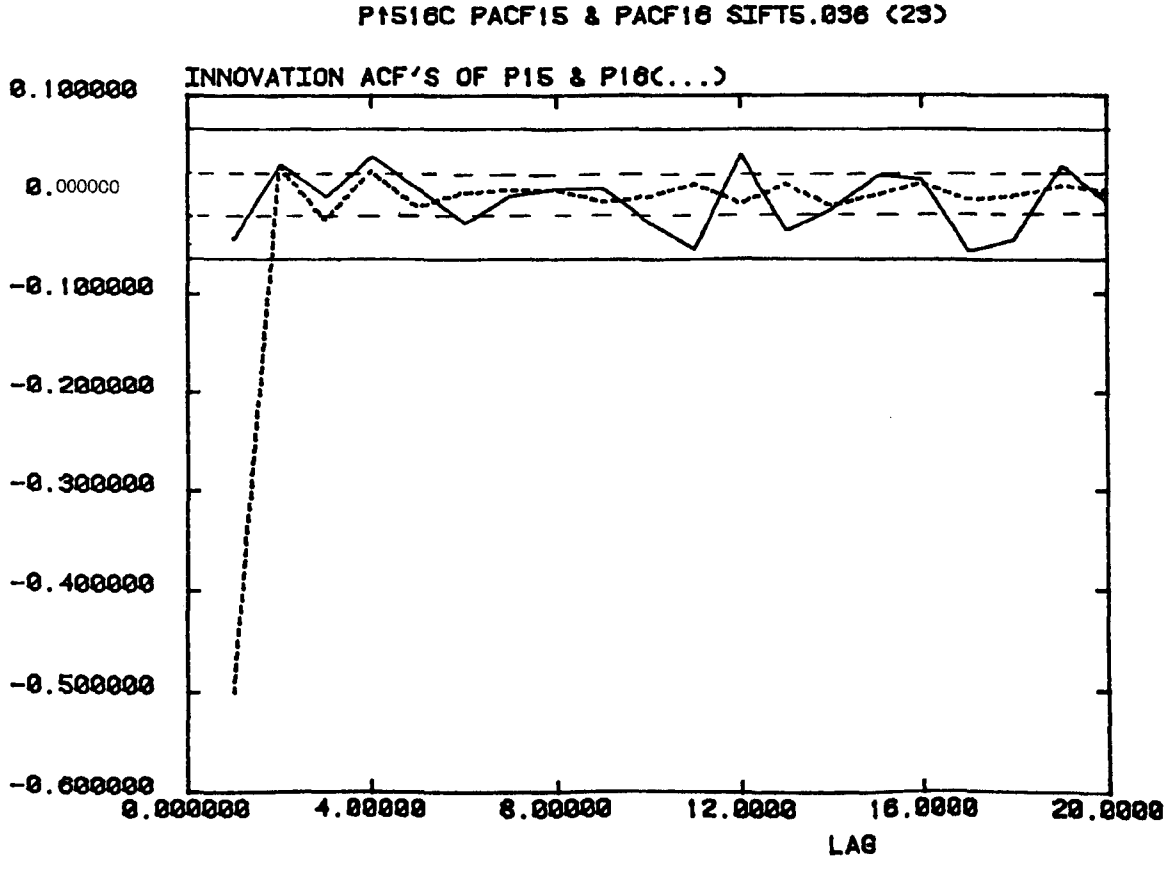
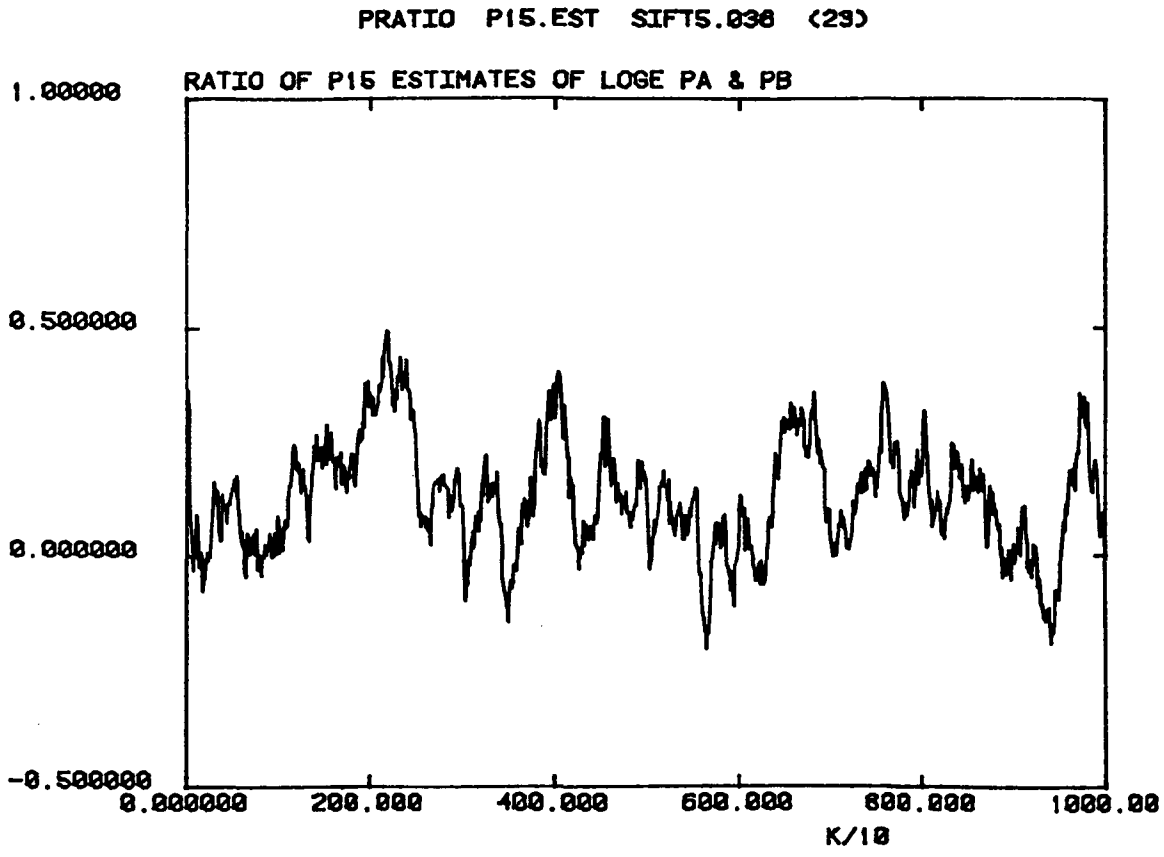
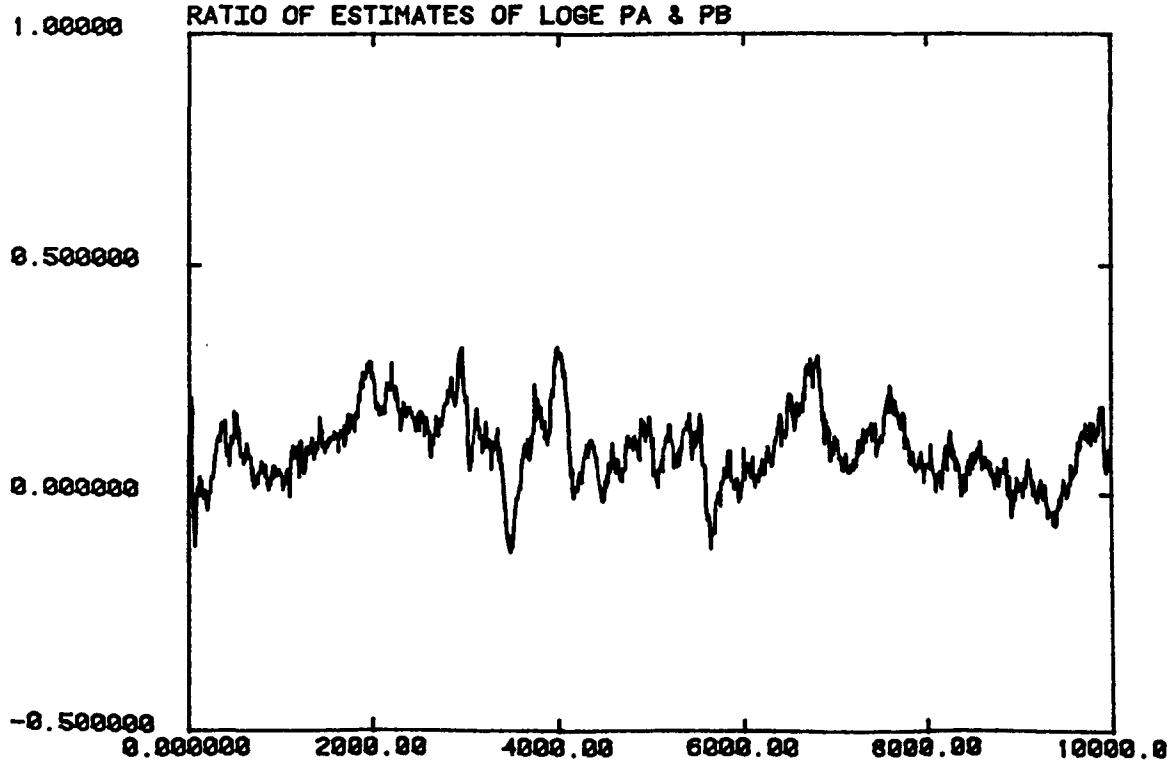


Figure 7.13 (k)



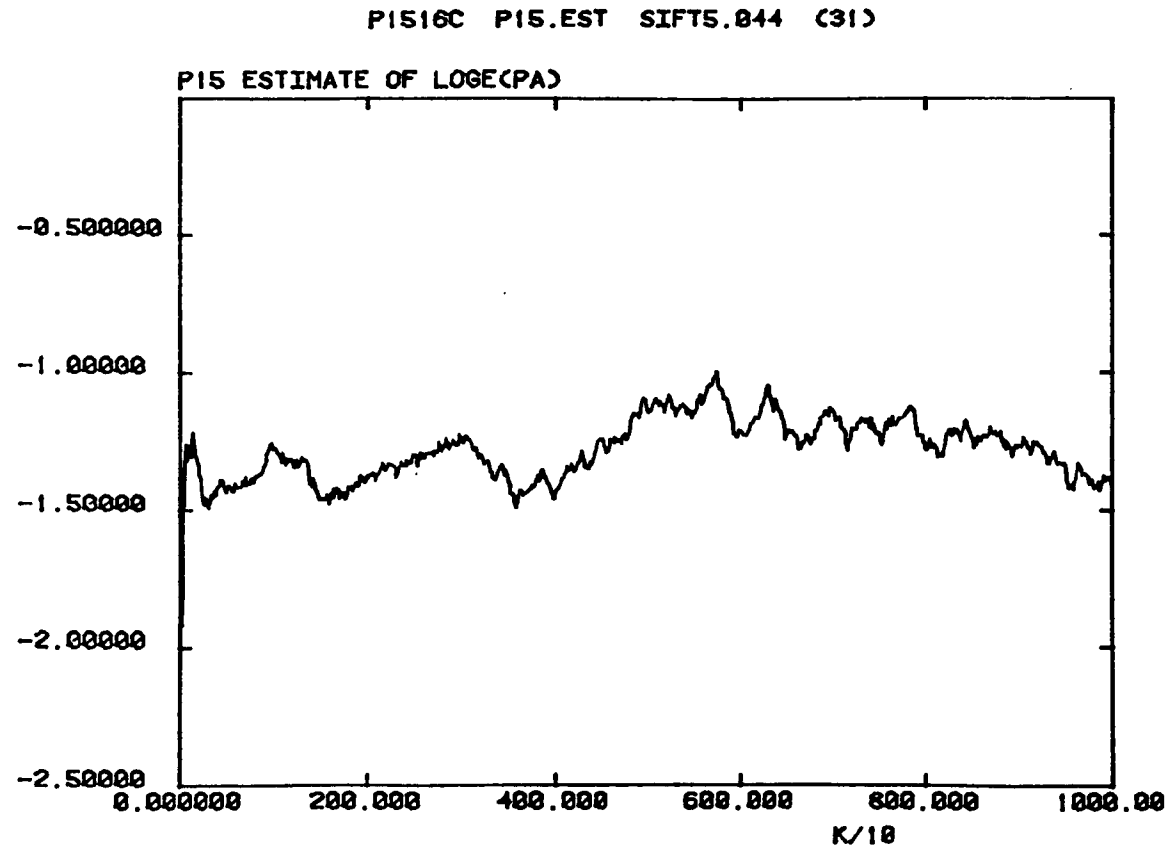
Ratio of P15 Estimates for Log (Normalised Power)

Figure 7.13 (l)



Ratio of P16 Estimates for Log (Normalised Power)

Figure 7.14 (a)



P15 Estimates of Log (Normalised Power) for SIFT 5·044:
Channel A (above)
Channel B (below)

Figure 7.14 (b)

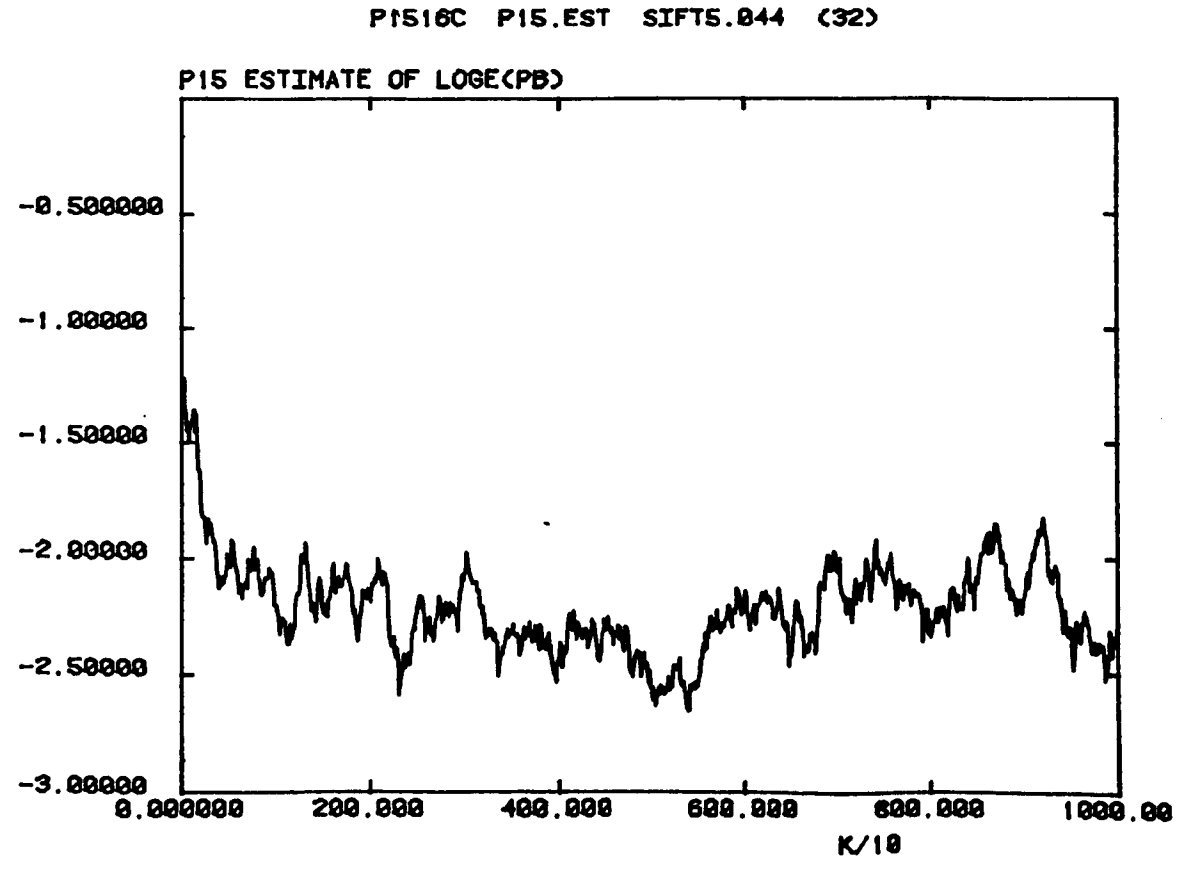
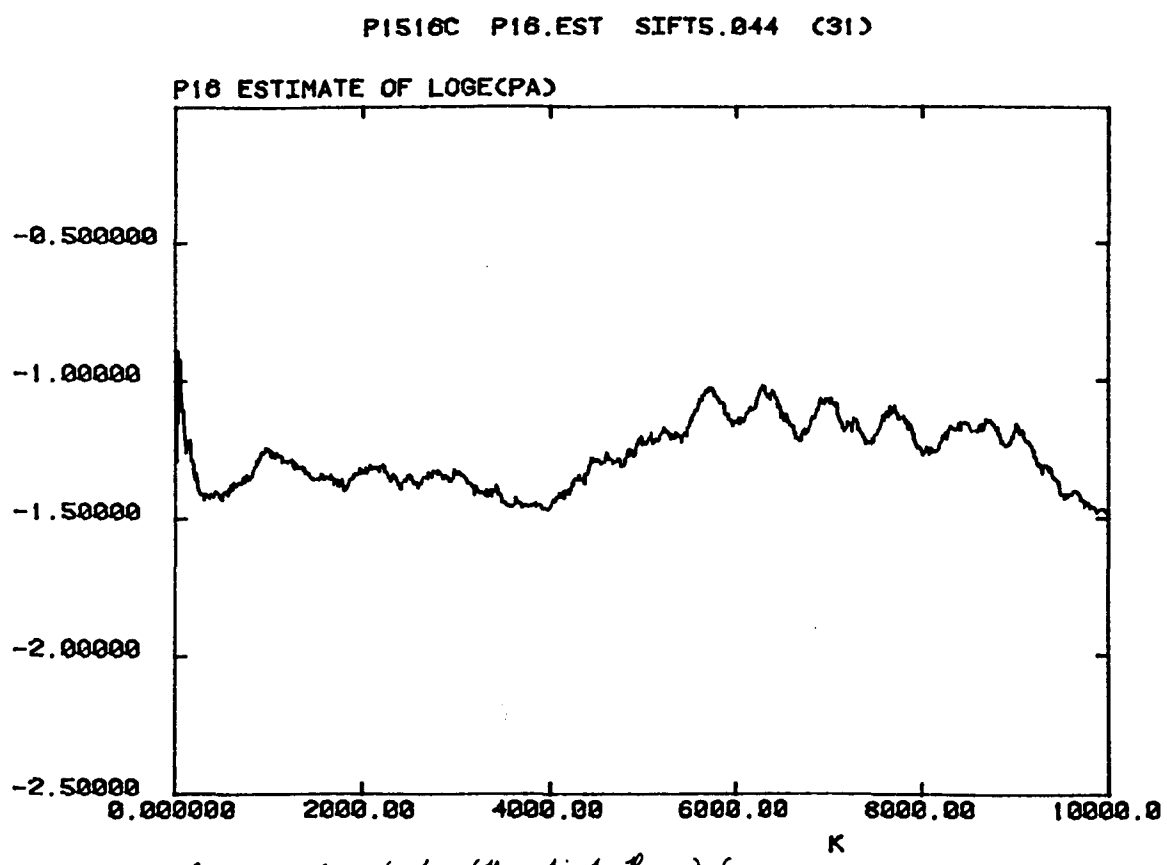


Figure 7.14 (c)



P16 Estimates of Log (Normalised Power) for SIFT 5·044:
Channel A (above)
Channel B (below)

Figure 7.14 (d)

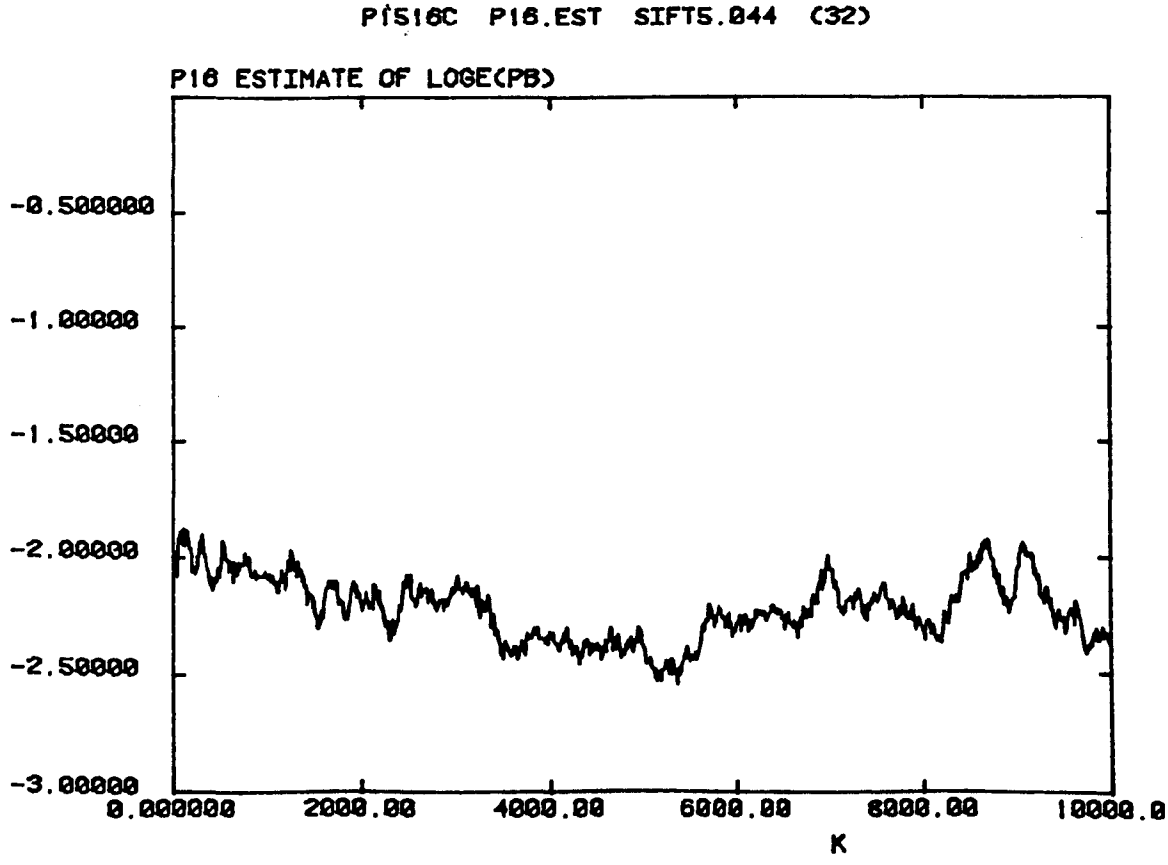
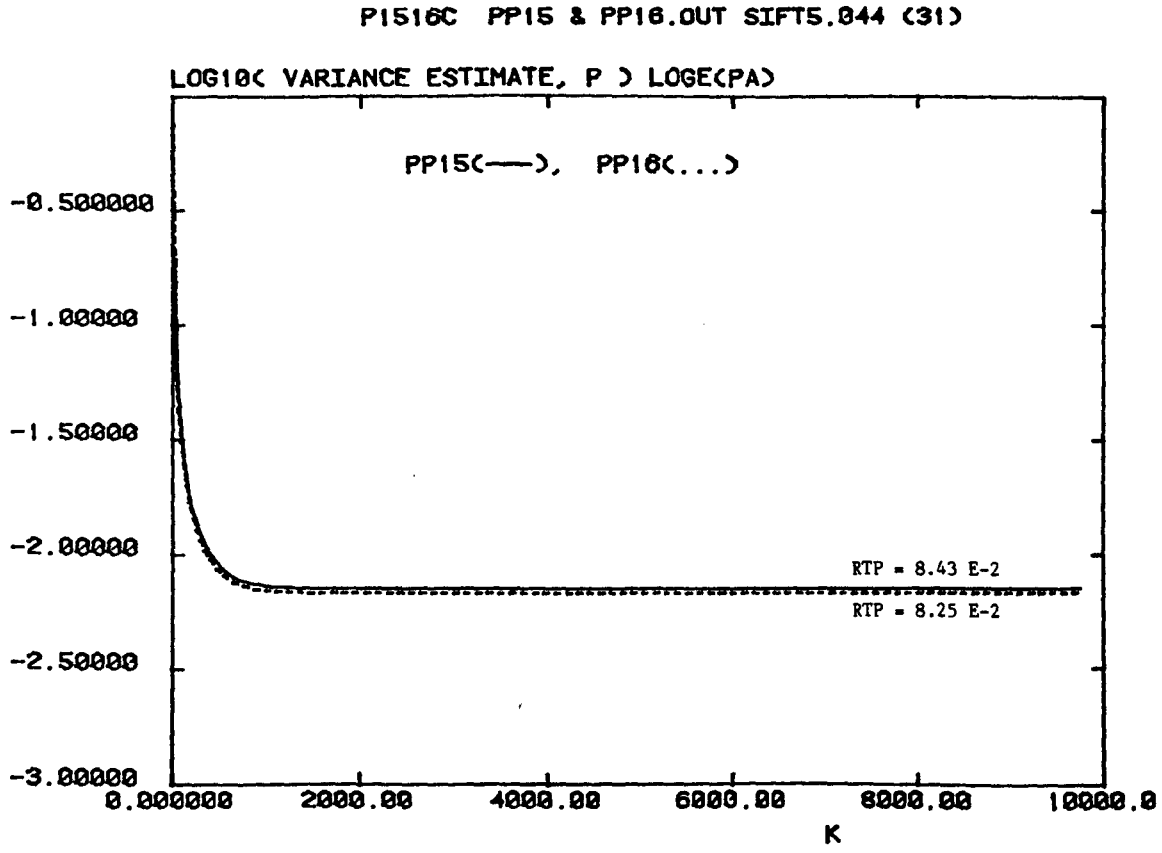


Figure 7.14 (e)



Error Covariance Estimates for SIFT 5·044: Channel A (above)
Channel B (below)

Figure 7.14 (f)

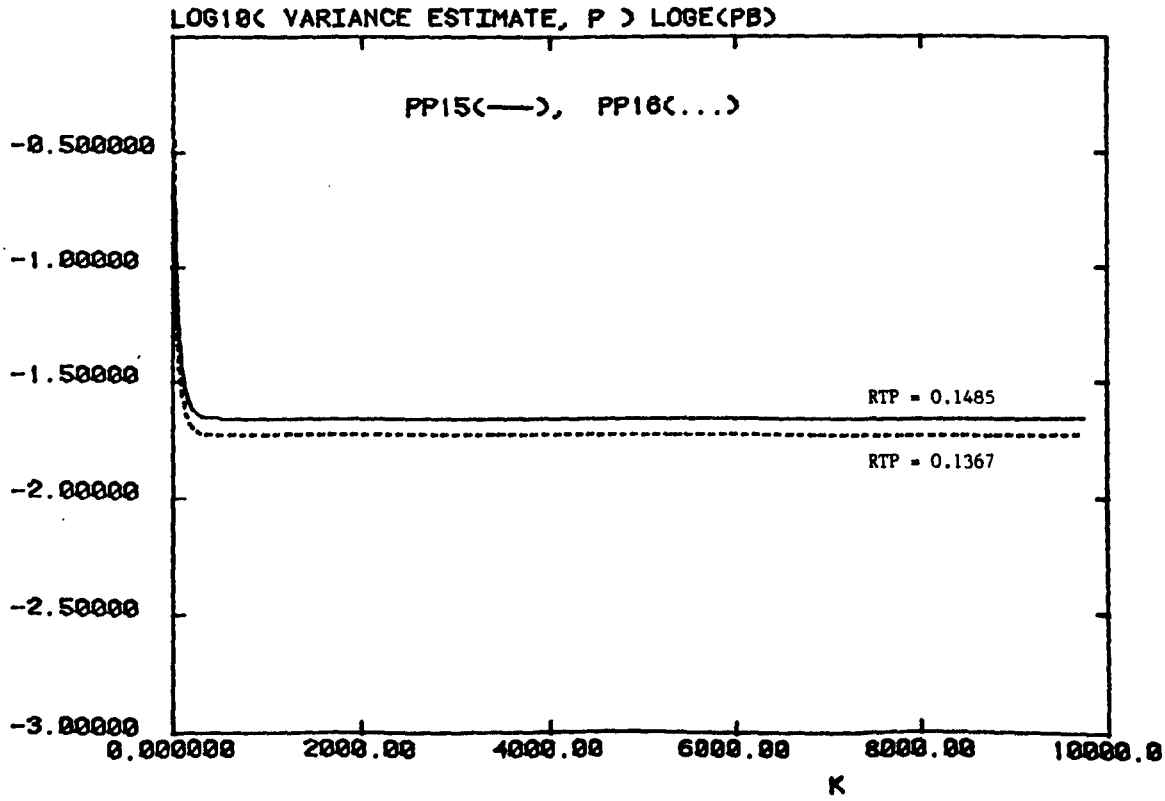
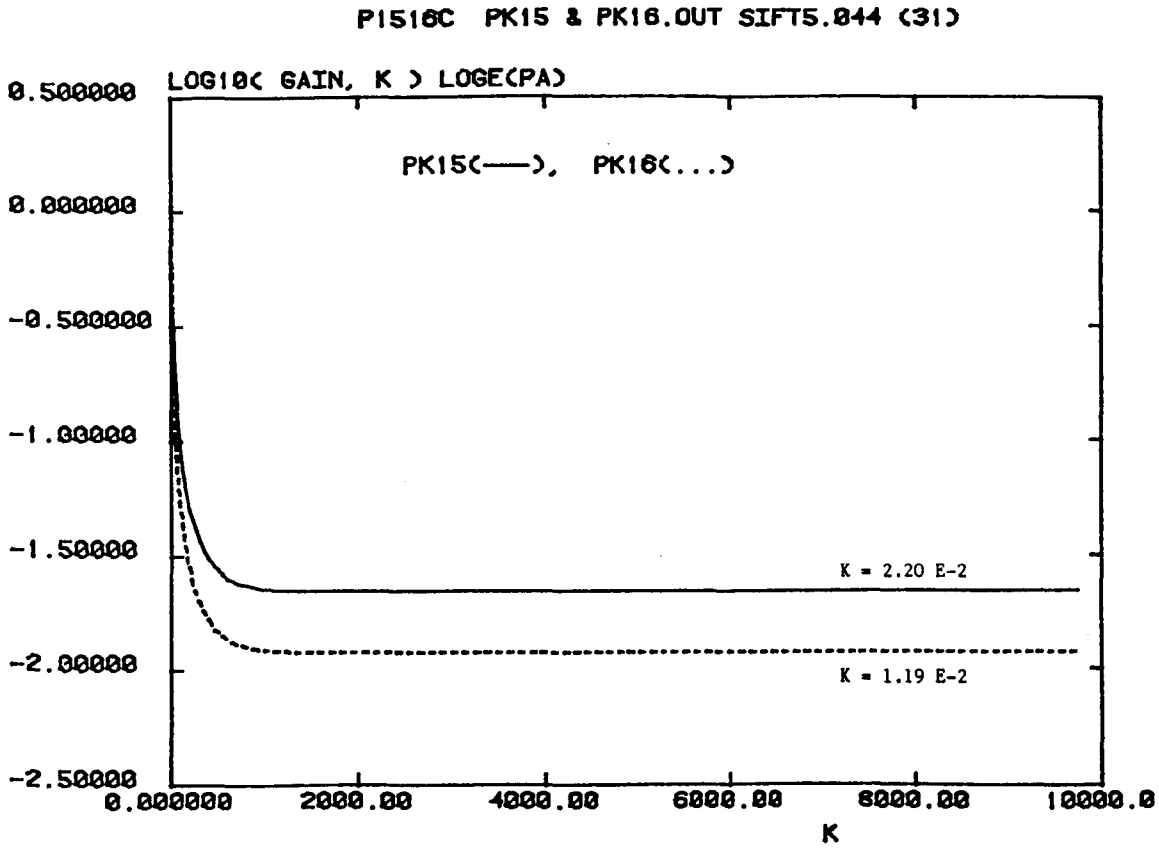


Figure 7.14 (g)



Kalman Gains used with SIFT 5·044: Channel A (above)
: Channel B (below).

Figure 7.14 (h)

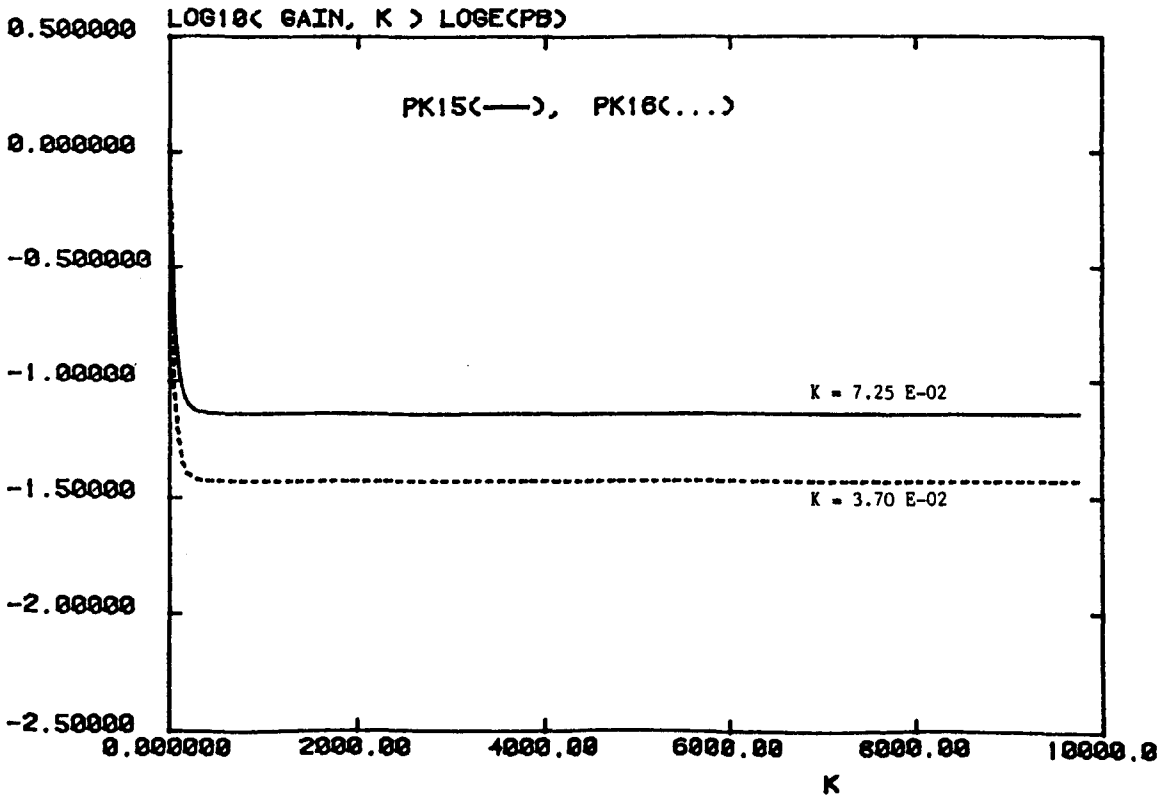
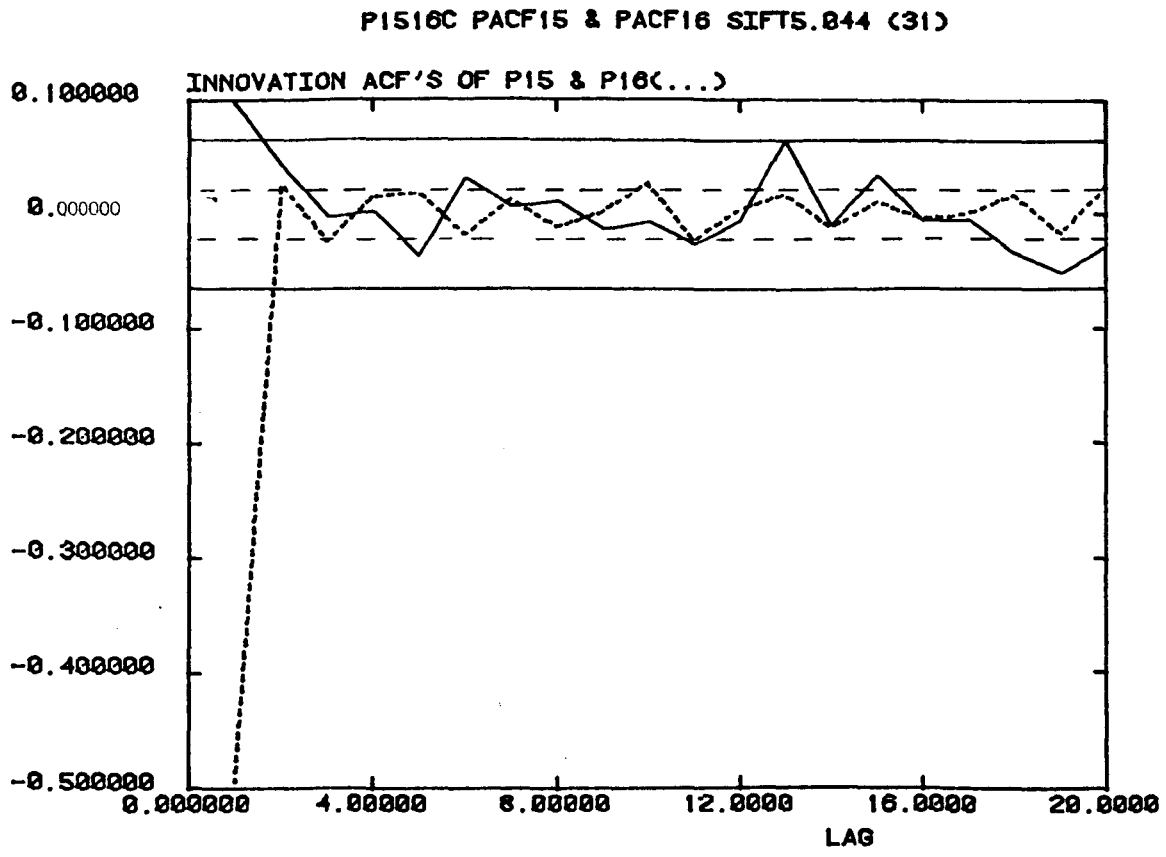


Figure 7.14 (i)



Innovation Sequence acfs for SIFT 5·044 : Channel A (above)
: Channel B (below)

Figure 7.14 (j)

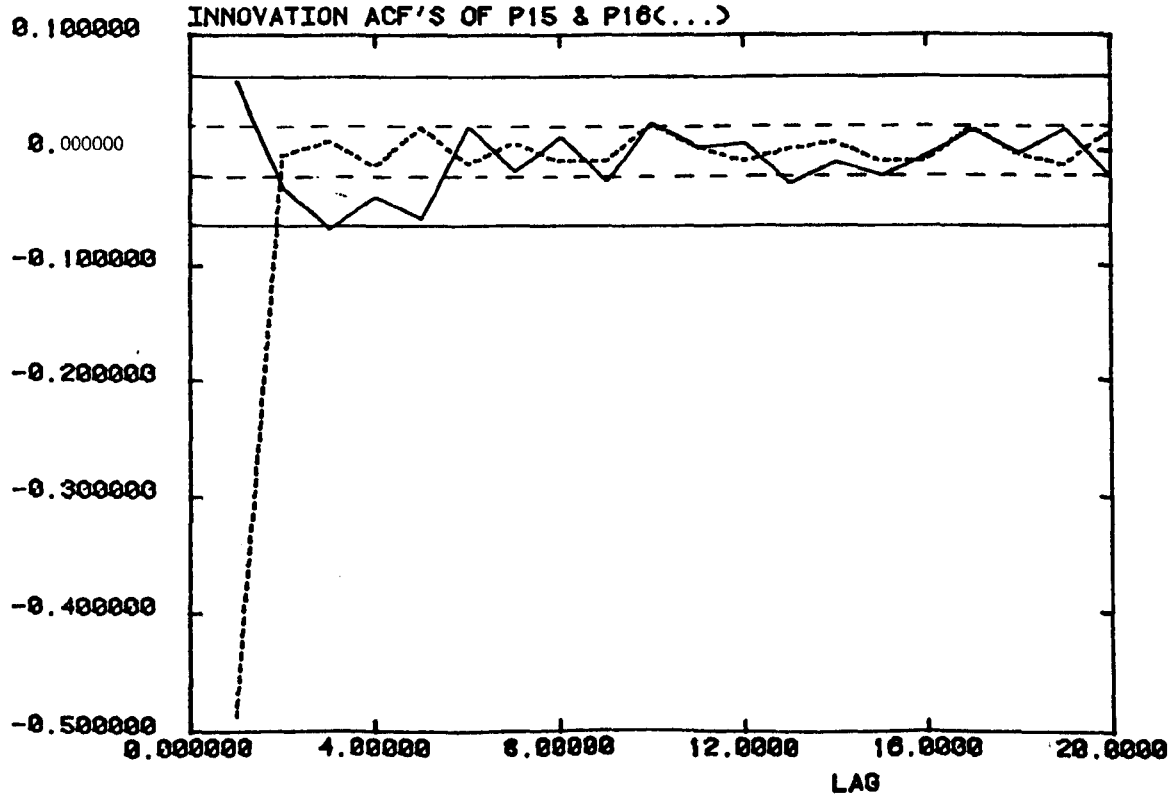
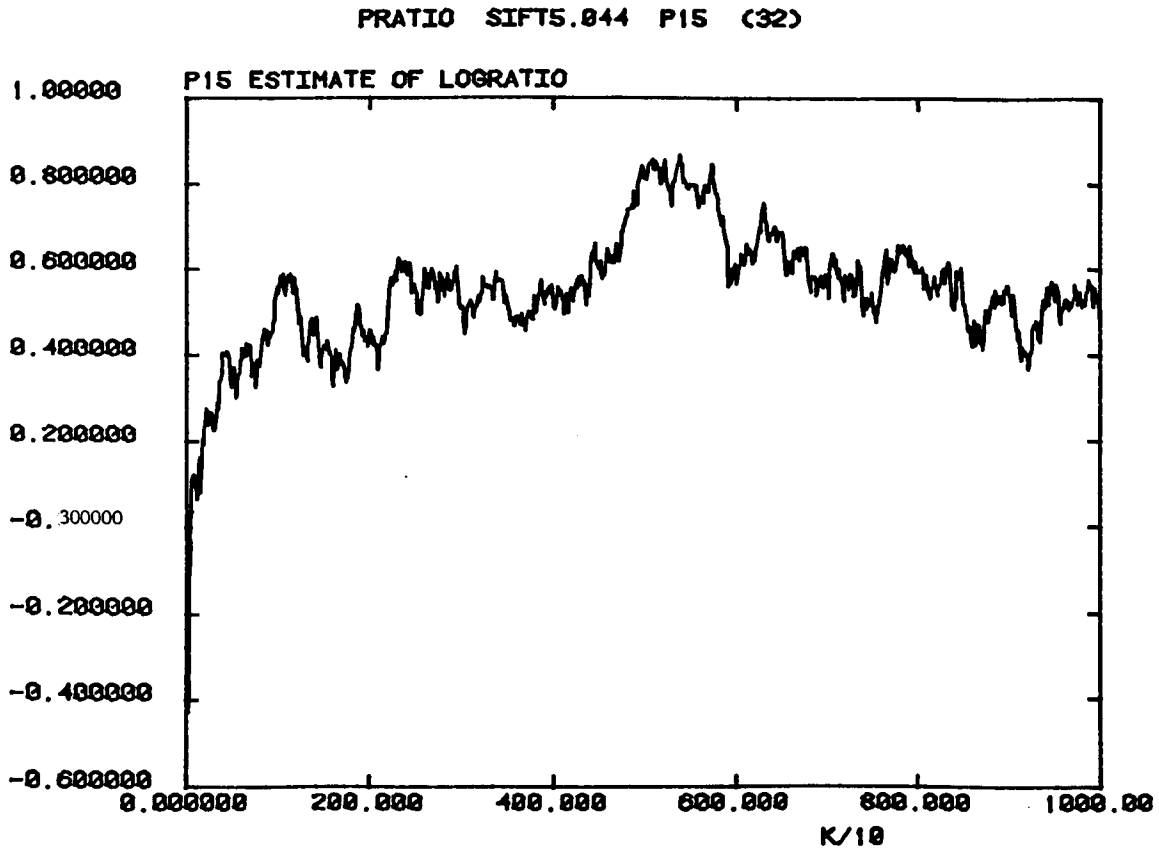


Figure 7.14 (k)



Estimates of the log Ratio for SIFT 5·044 : P15 (above)
: P16 (below)

Figure 7.14 (l)

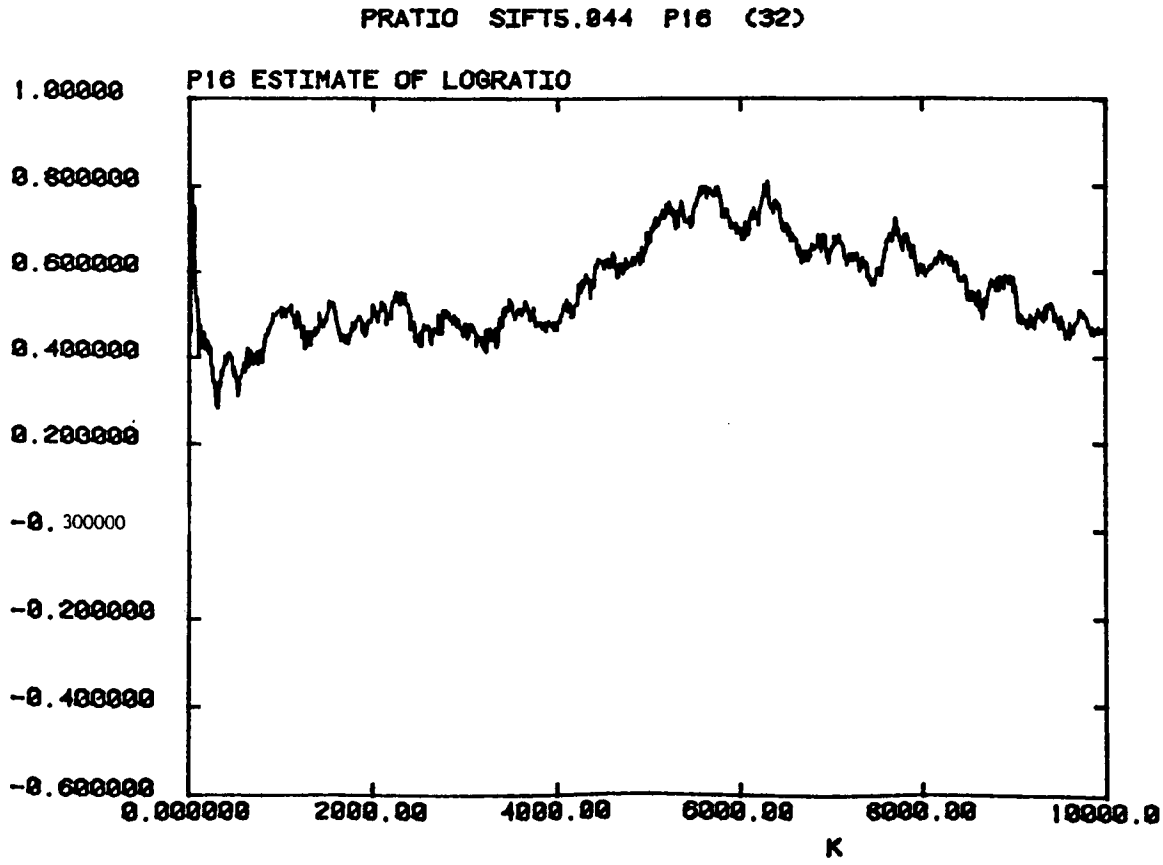
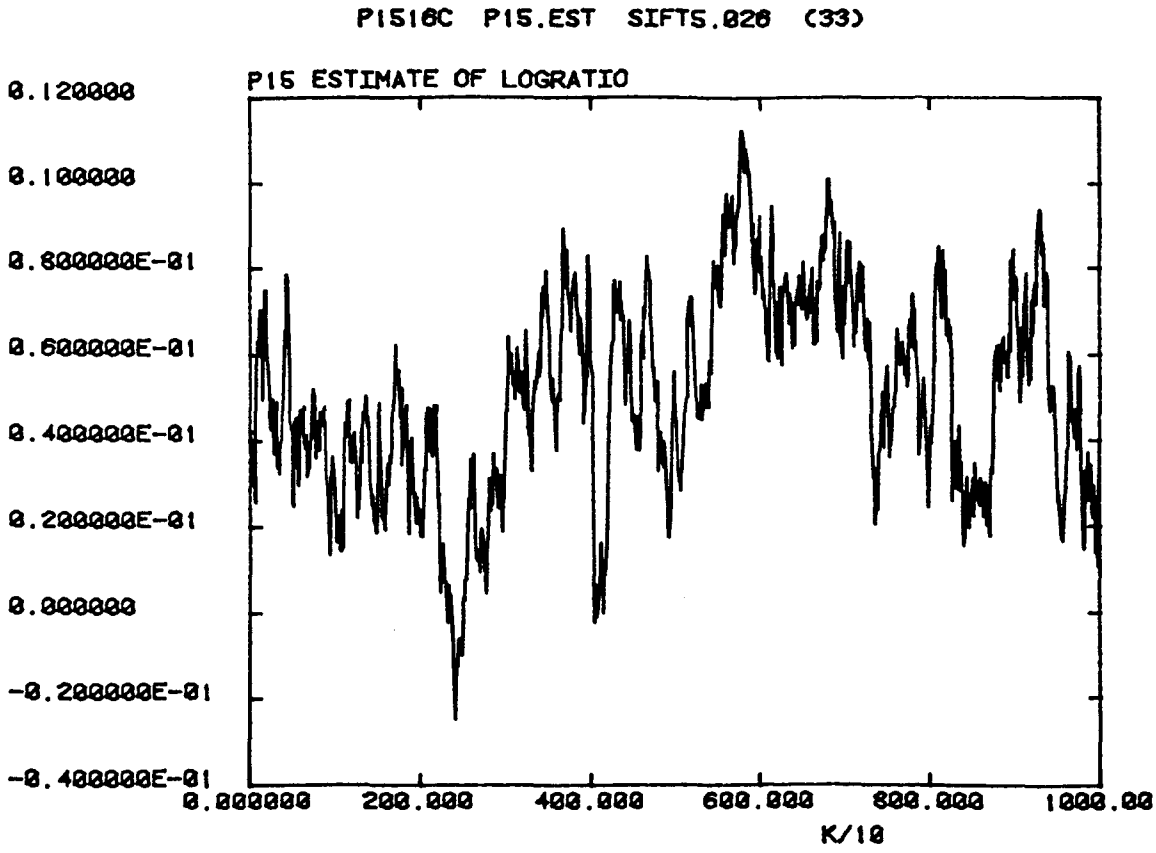


Figure 7.15 (a)



Estimates of the log Ratio for SIFT5·026: P15 (above)
P16 (below)

Figure 7.15 (b)

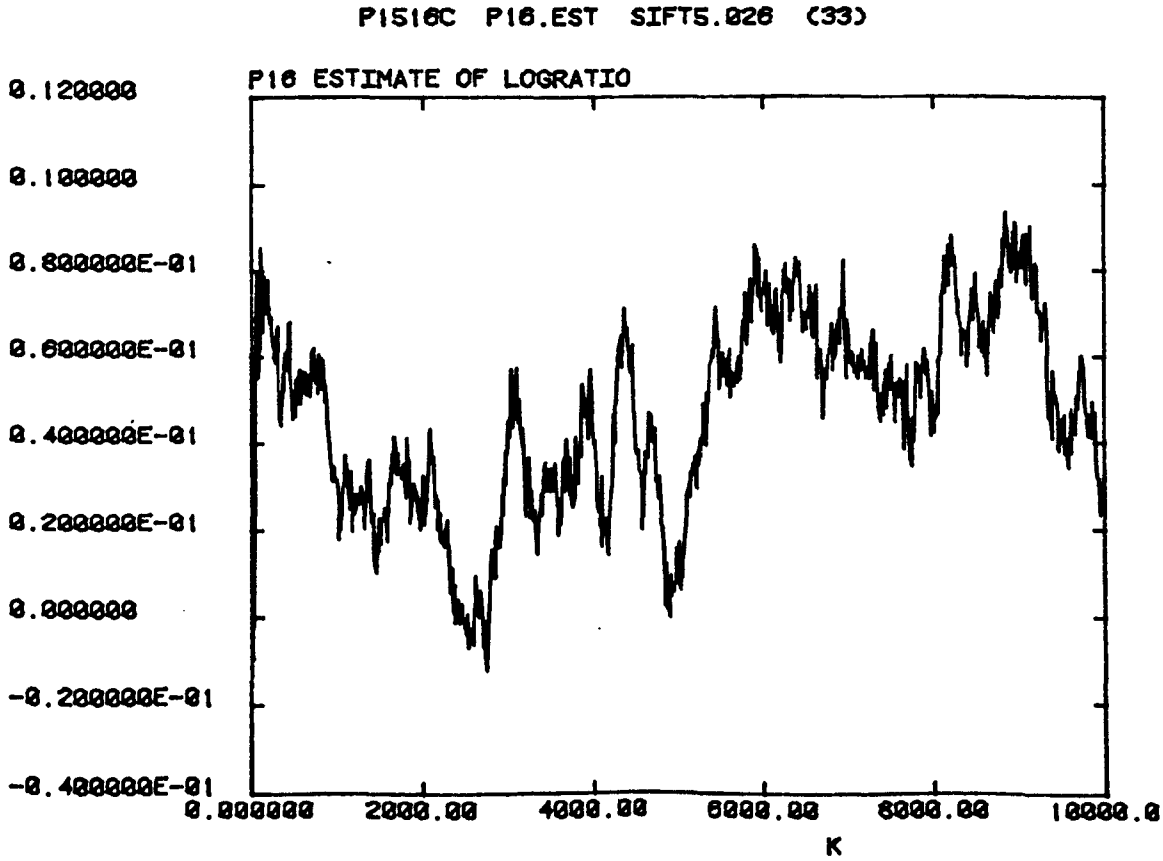
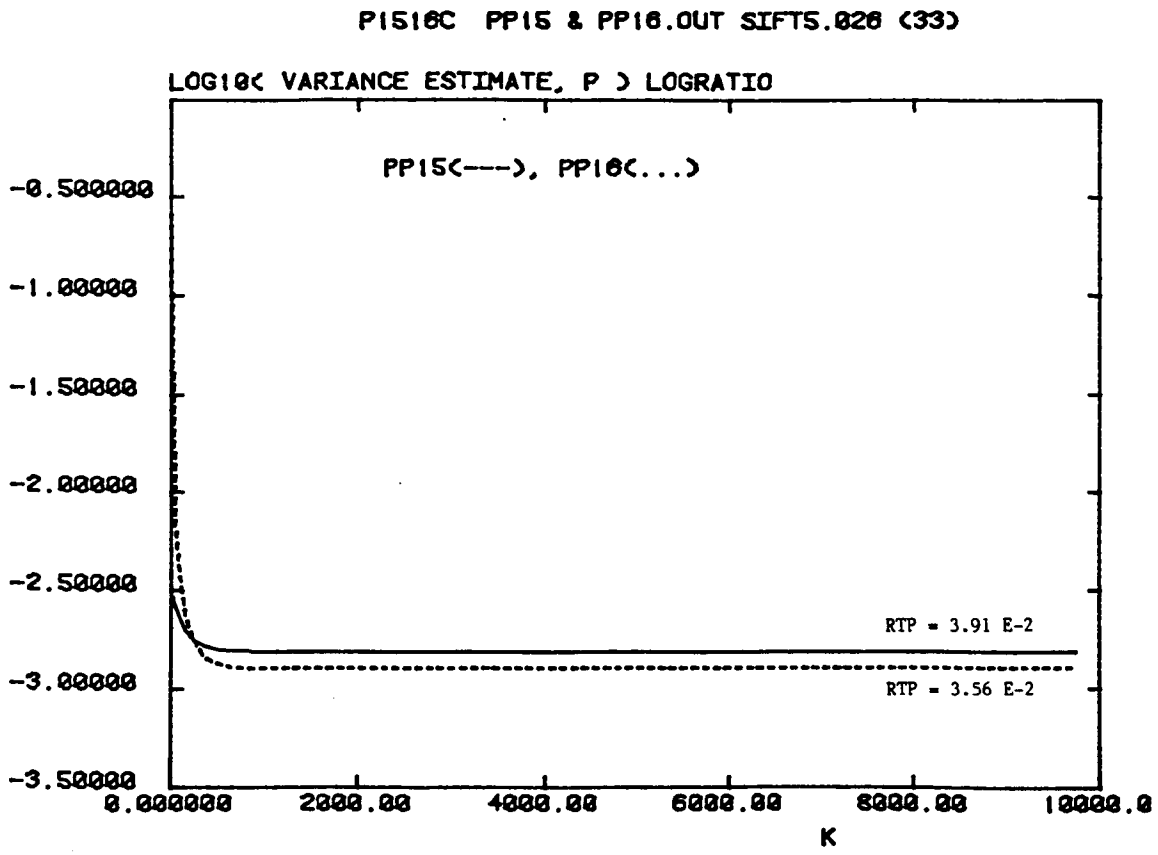
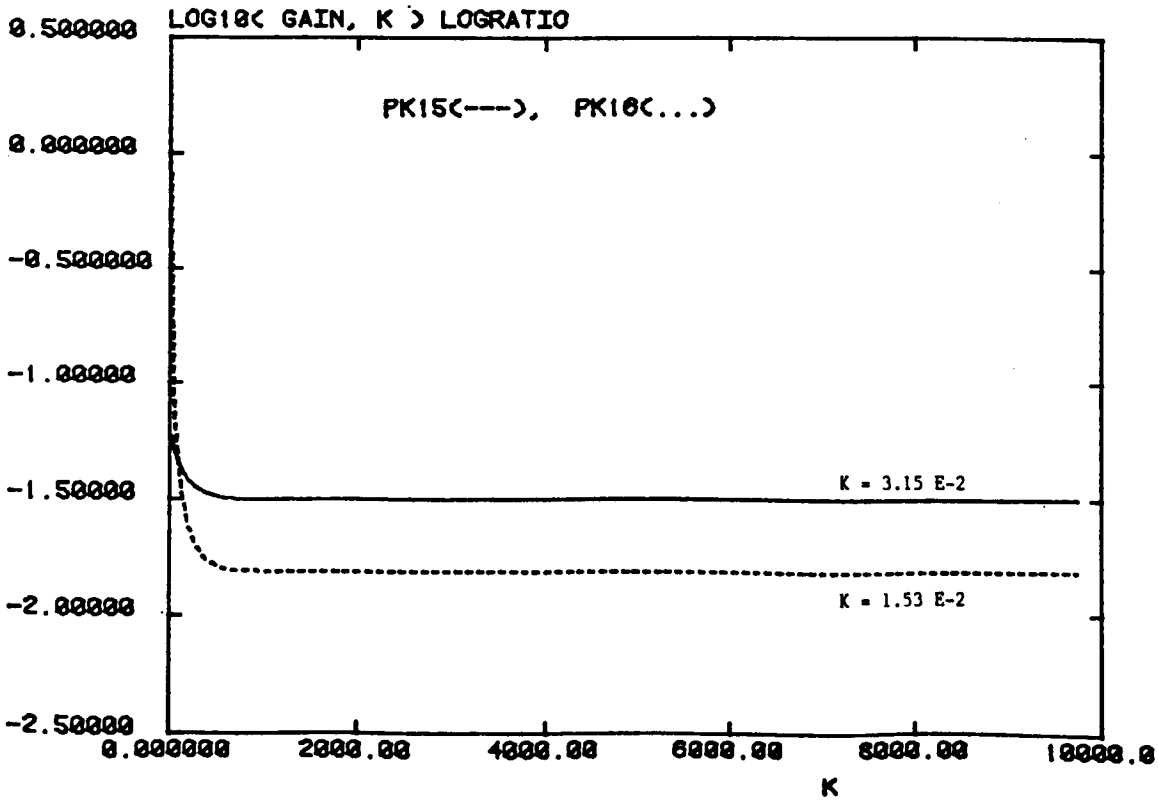


Figure 7.15 (c)



Error Covariance Estimates for SIFT 5·026 (above)
Kalman Gains used with SIFT 5·026 (below)

Figure 7.15 (d)



Innovation Sequence acfs for SIFT 5·026

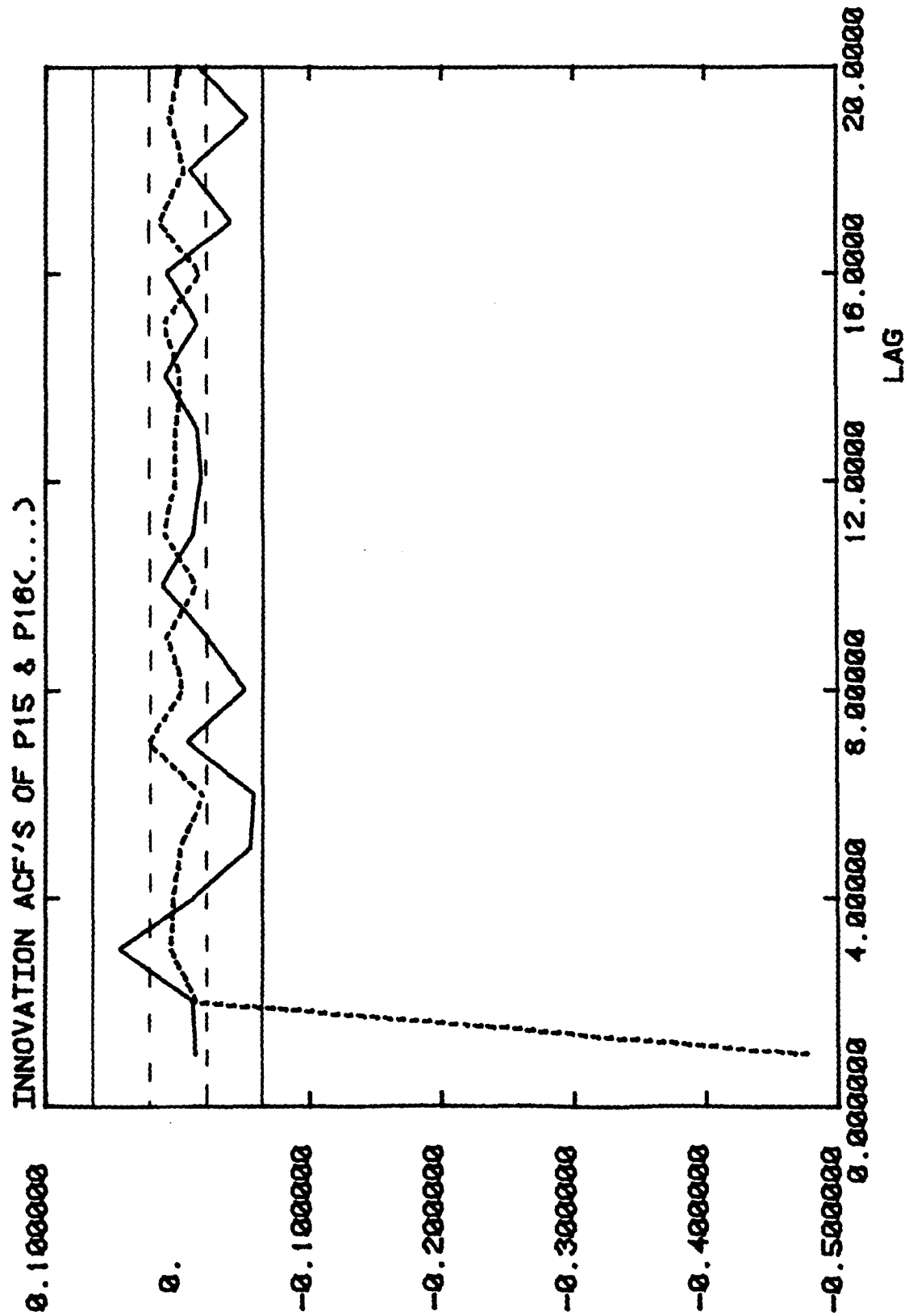
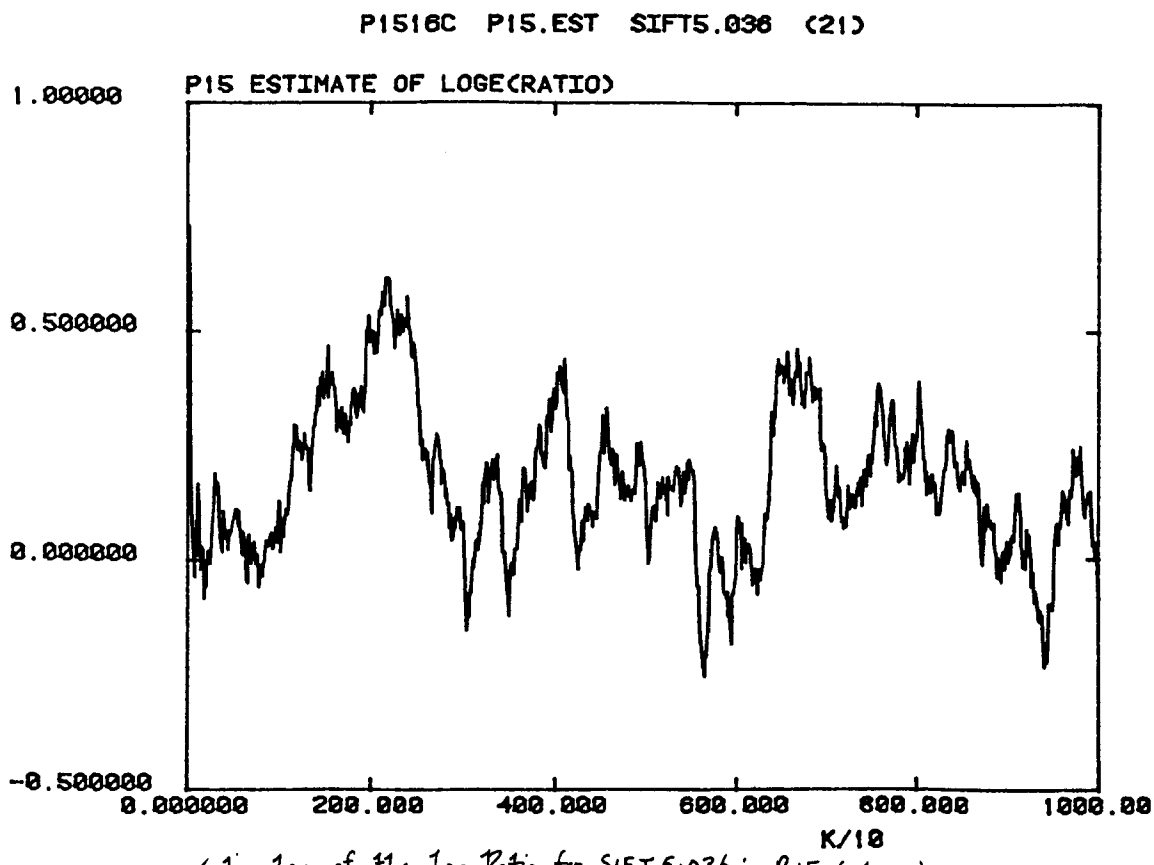


Figure 7.15 (e)

Figure 7.16 (a)



Estimates of the Log Ratio for SIFT 5·036 : p_{15} (above)

: p_{16} (below)

Figure 7.16 (b)

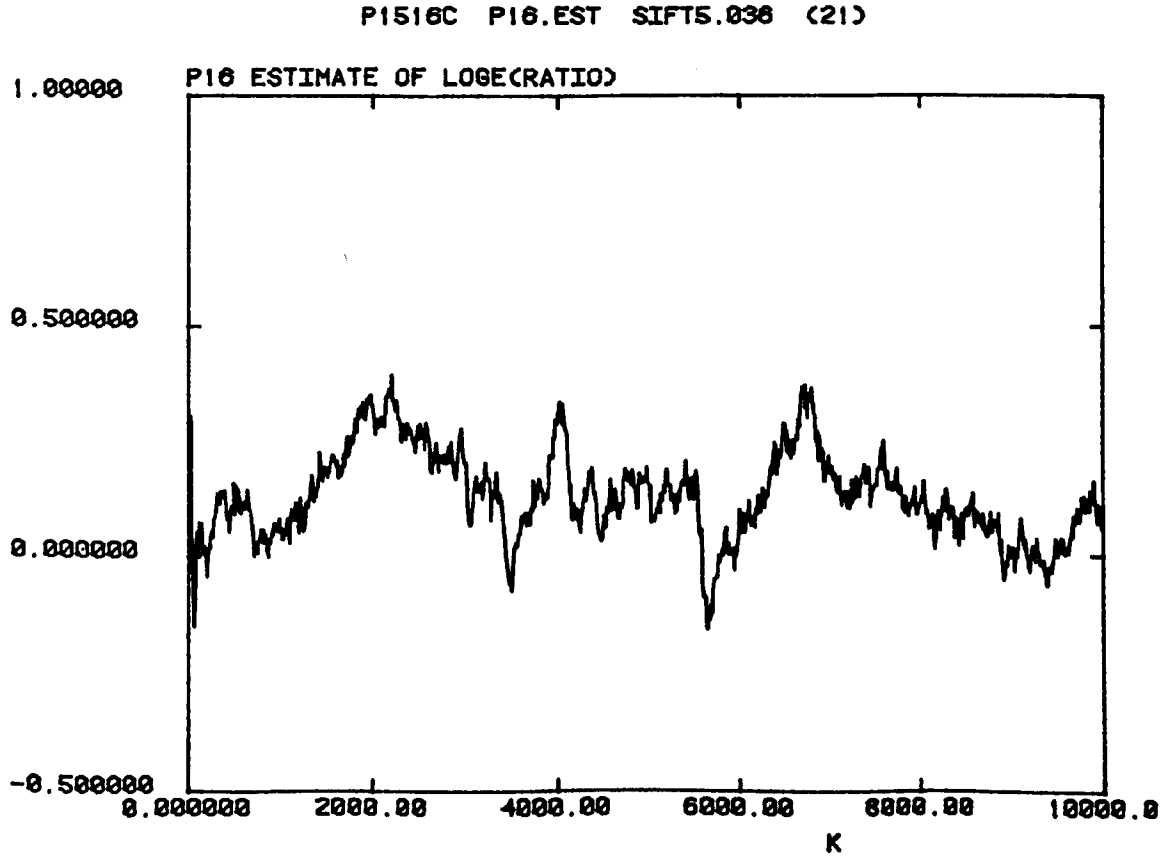
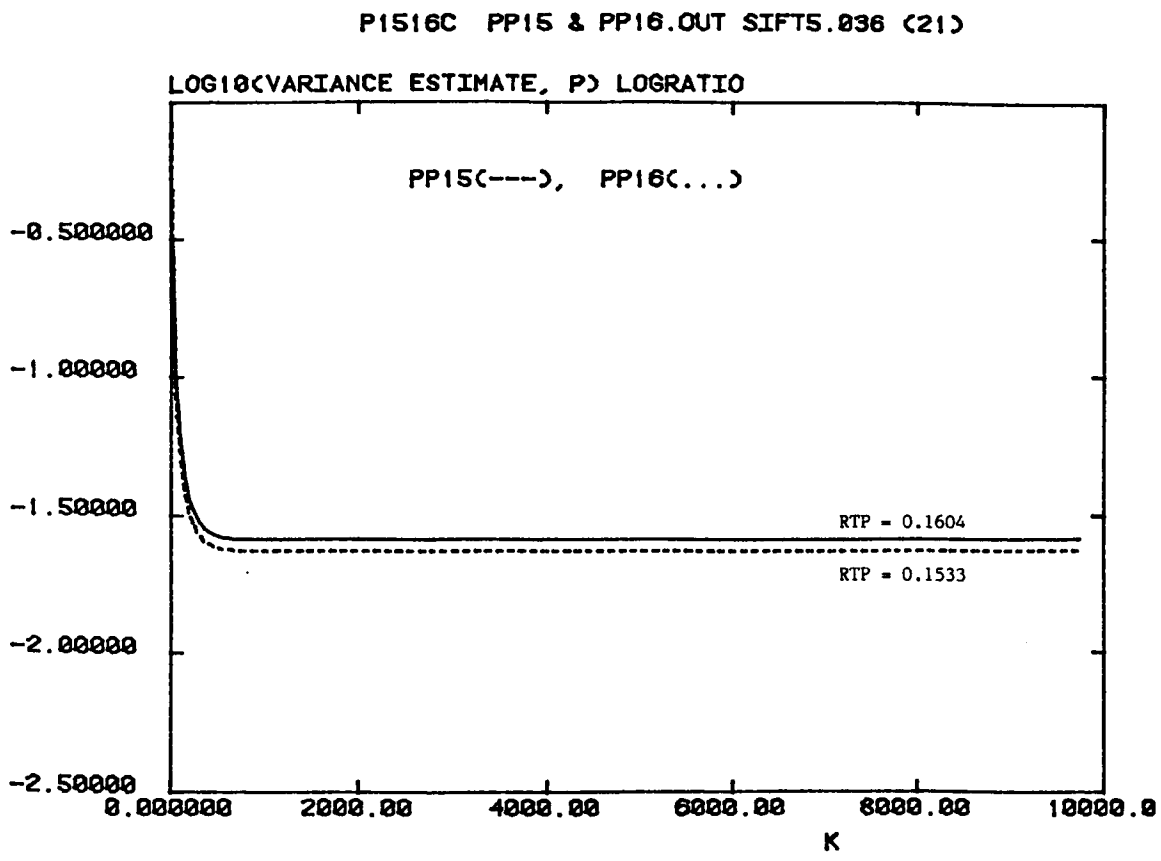


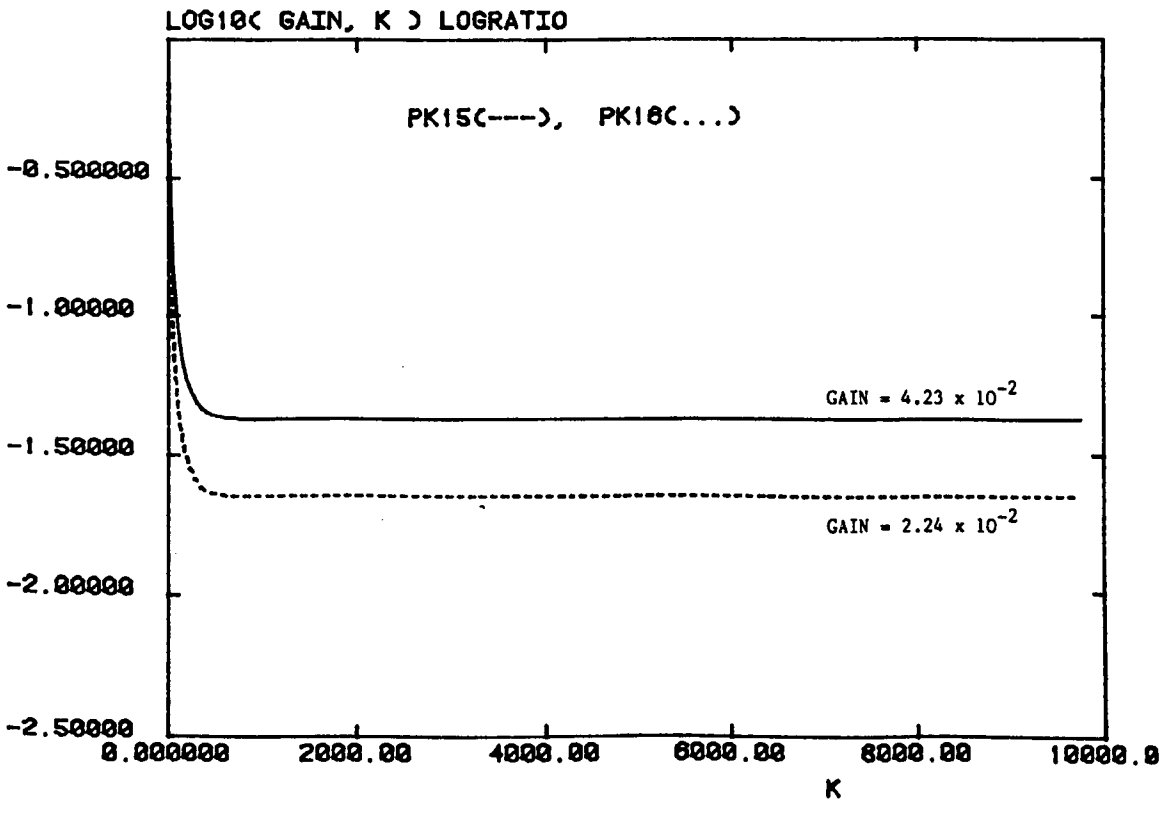
Figure 7.16 (c)



Error Covariance Estimates for SIFT 5.036 (above)

Kalman Gains Used With SIFT 5.036 (below)

Figure 7.16 (d)



Innovation Sequence acfs for SIFT 5·036

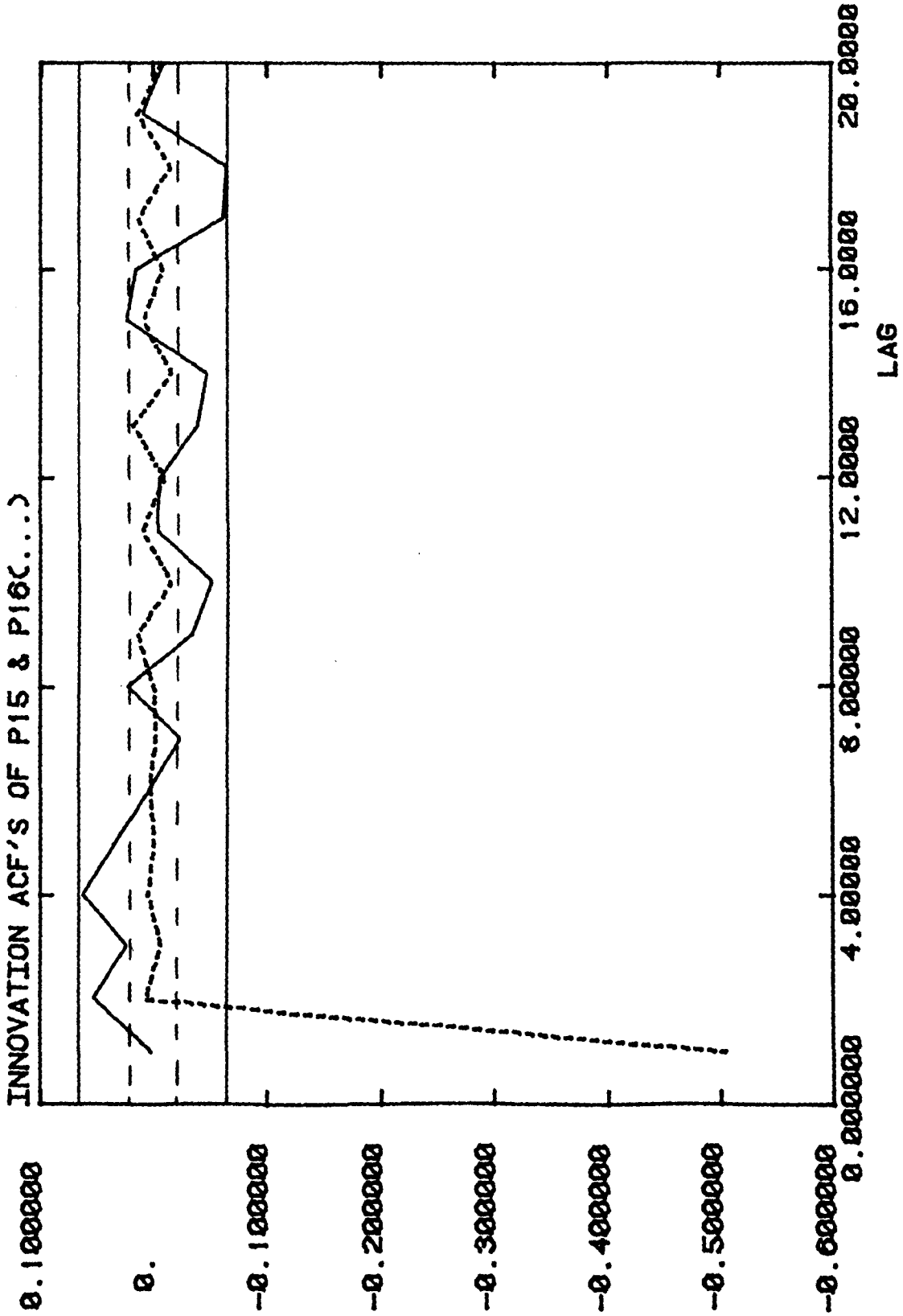
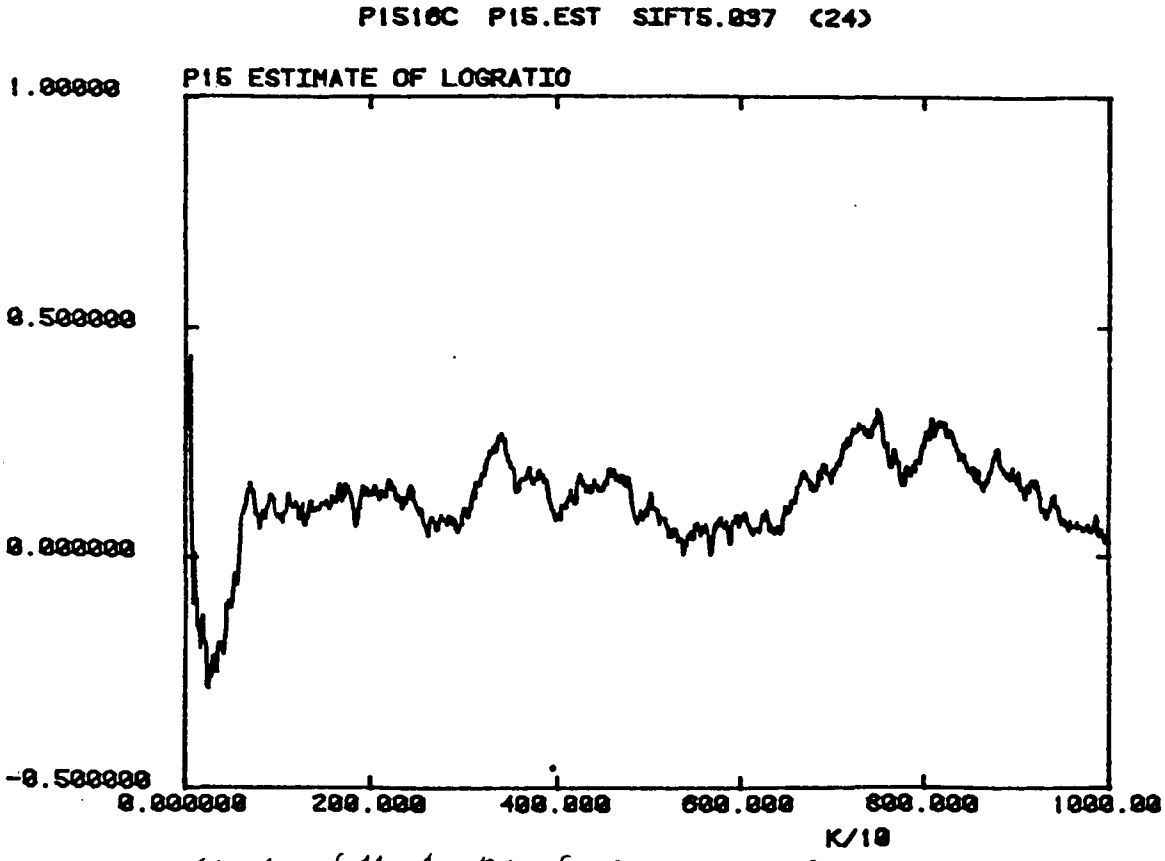


Figure 7.16 (e)

Figure 7.17 (a)



Estimates of the log Ratio for SIFT 5·037 : P15 (above)

: P16 (below)

Figure 7.17 (b)

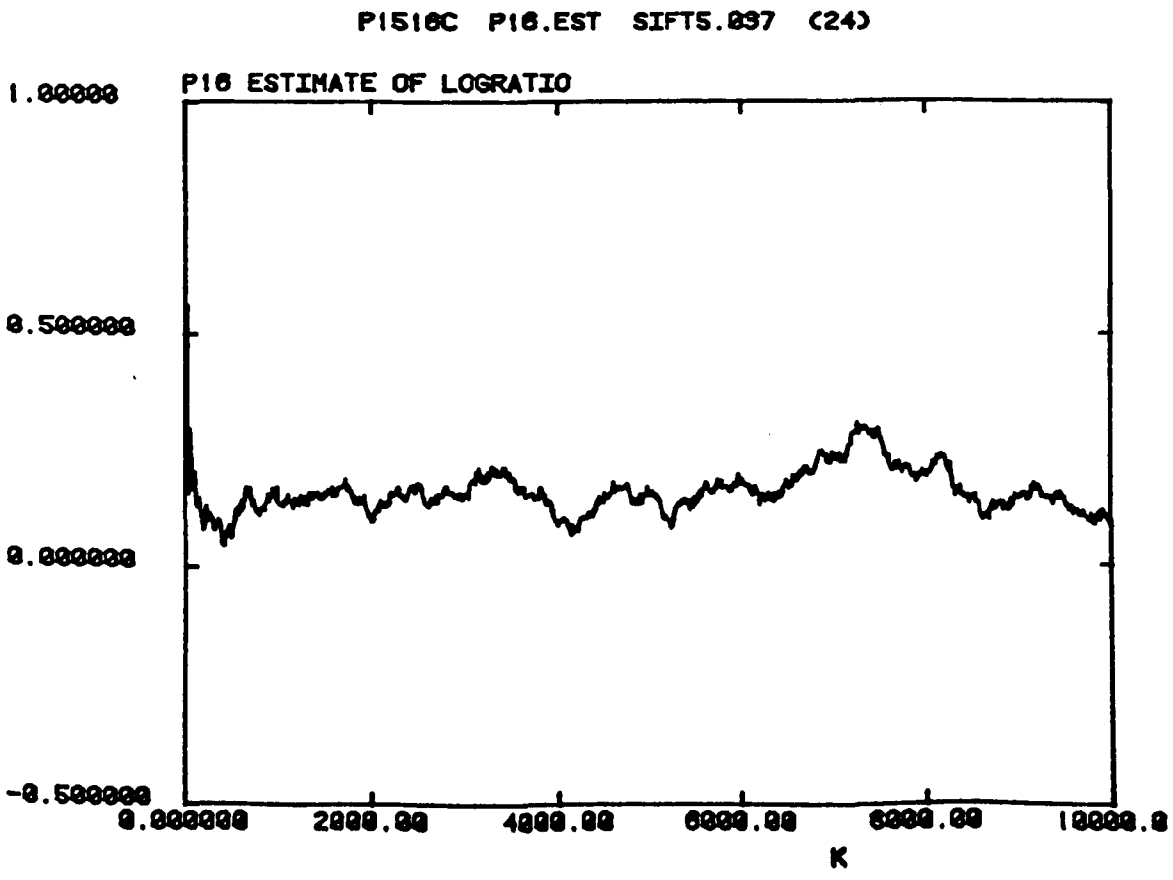
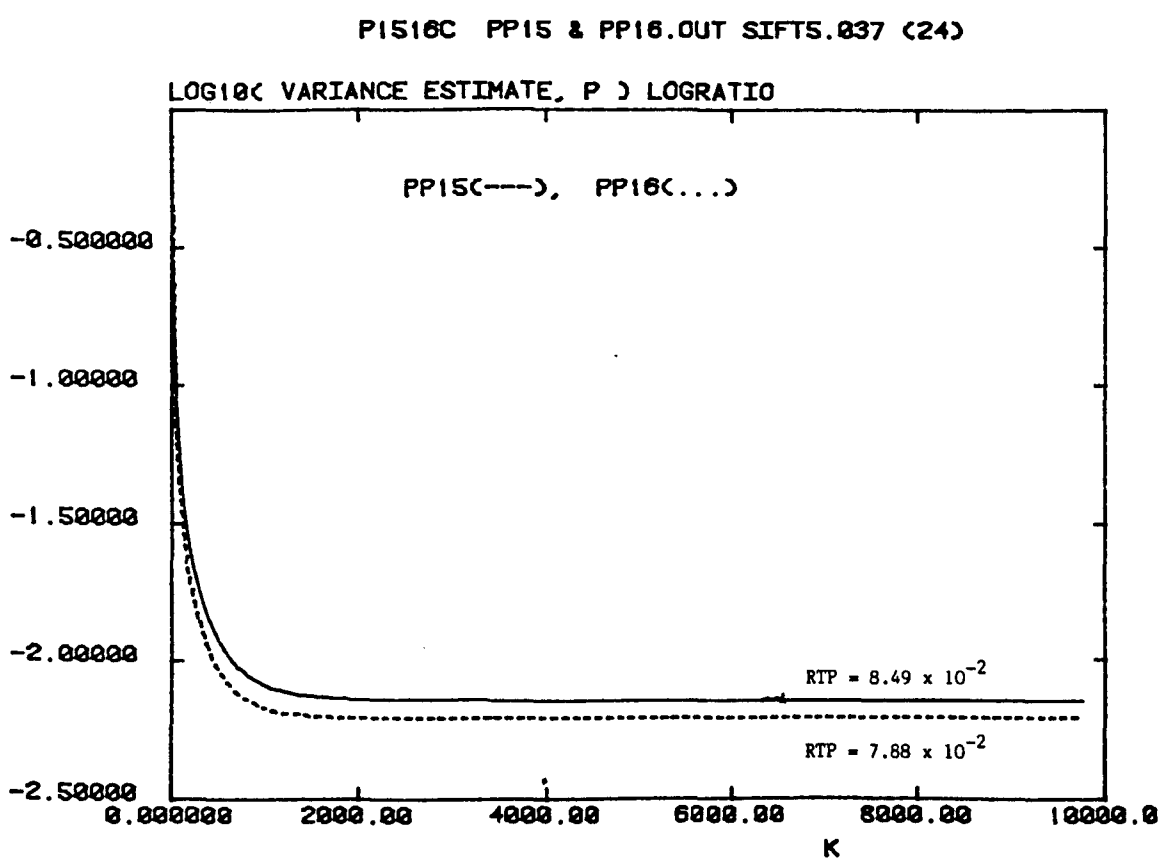


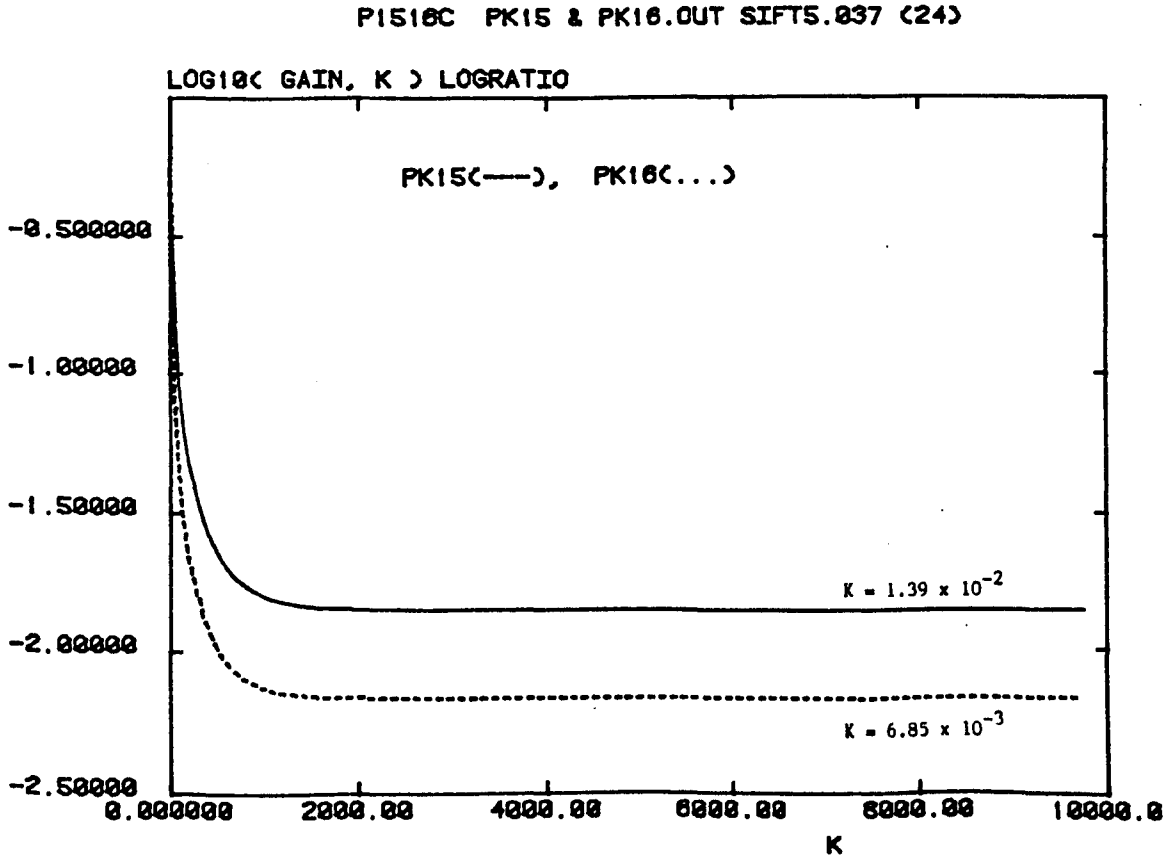
Figure 7.17 (c)



Error Covariance Estimates for SIFT5·037 (above)

Kalman Gains Used With SIFT5·037 (below)

Figure 7.17 (d)



Innovation Sequence acfs for SIFT5·037

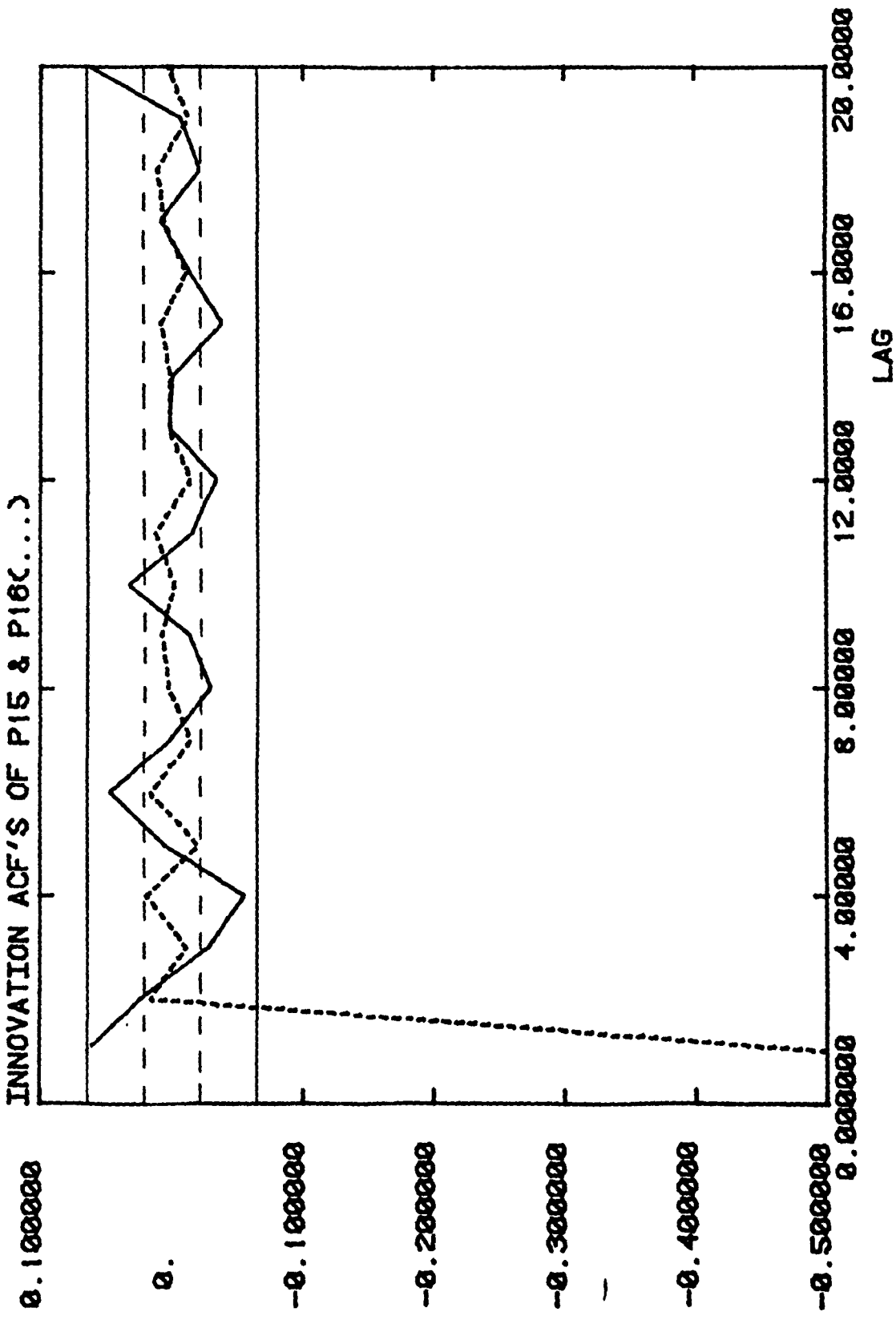
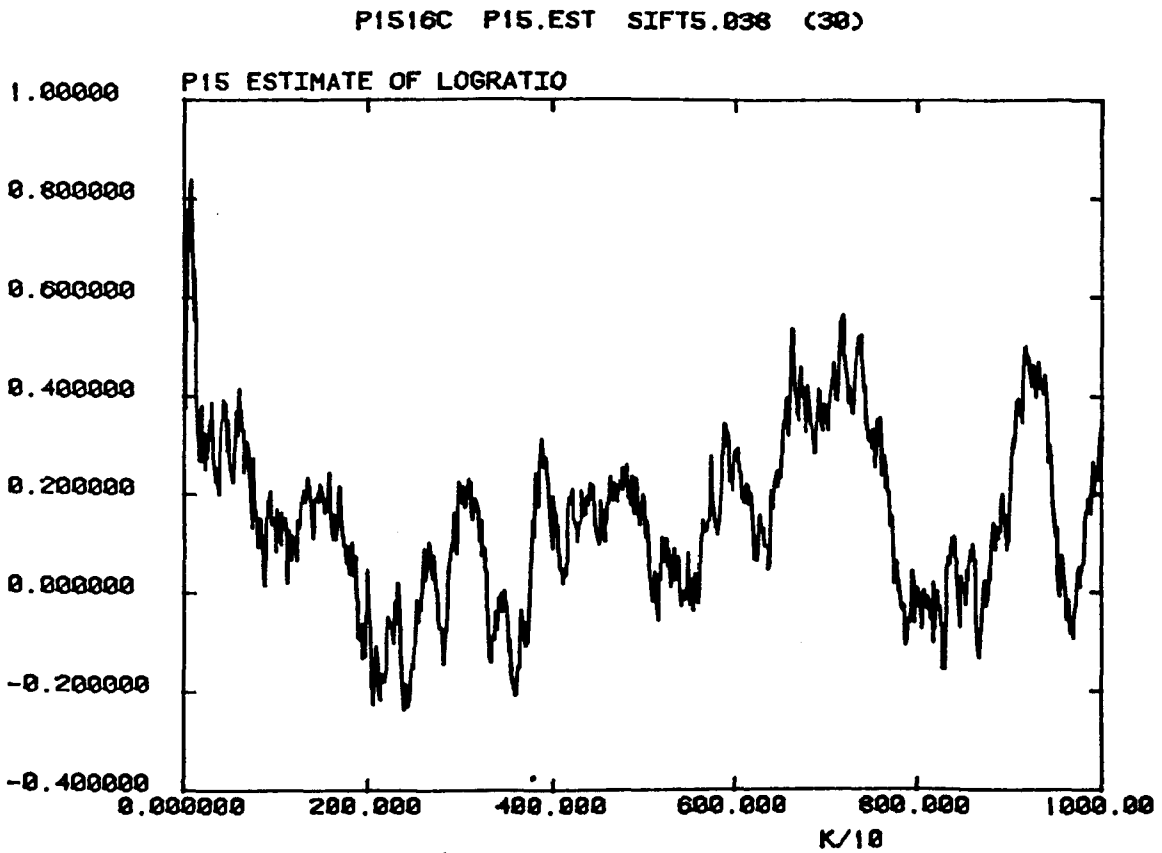


Figure 7.17 (e)

Figure 7.18 (a)



Estimates of the log Ratio for SIFT 5·038 : P15 (above)

: P16 (below)

Figure 7.18 (b)

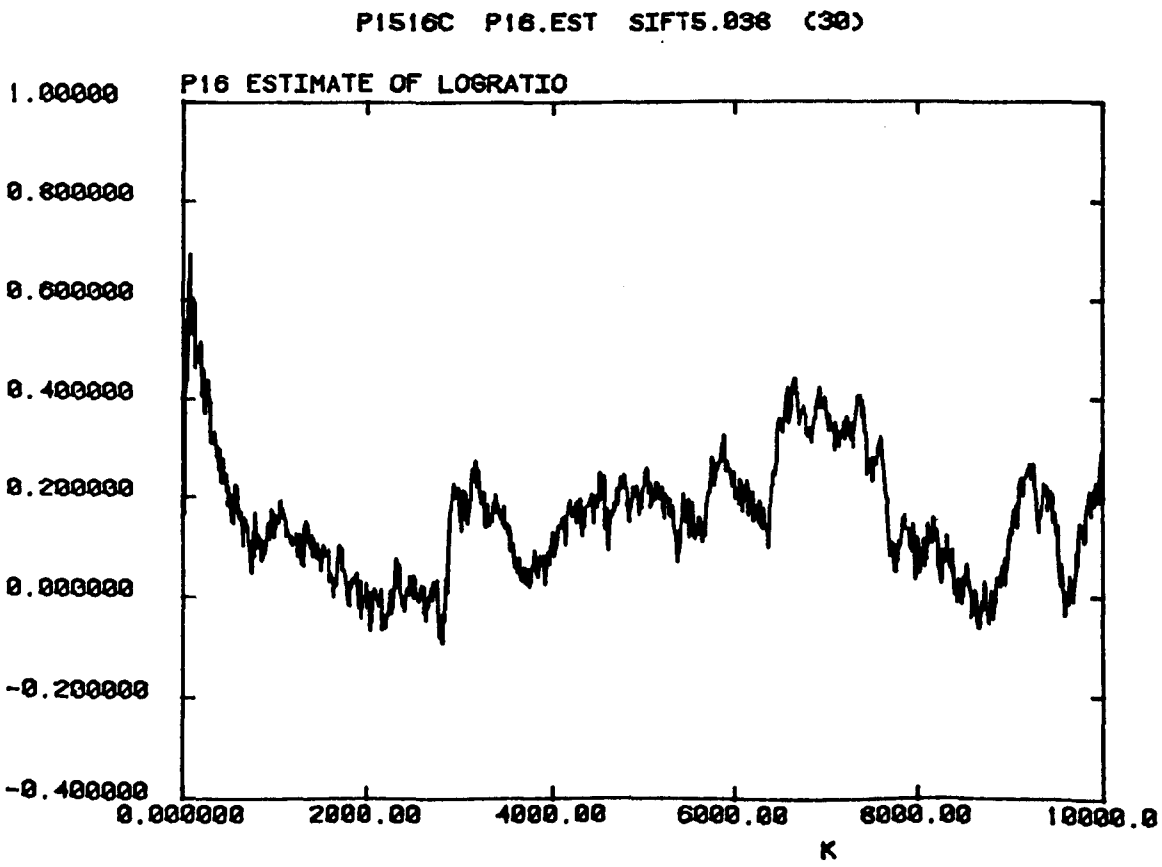
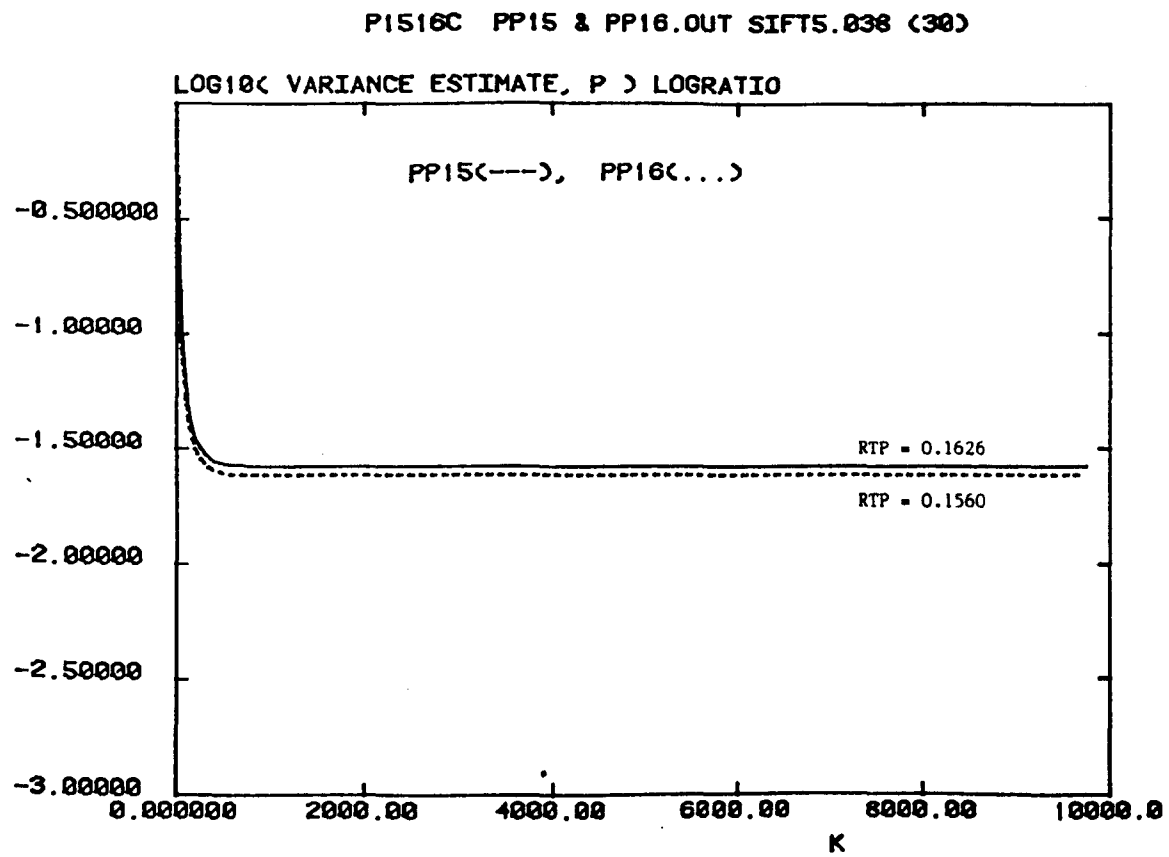


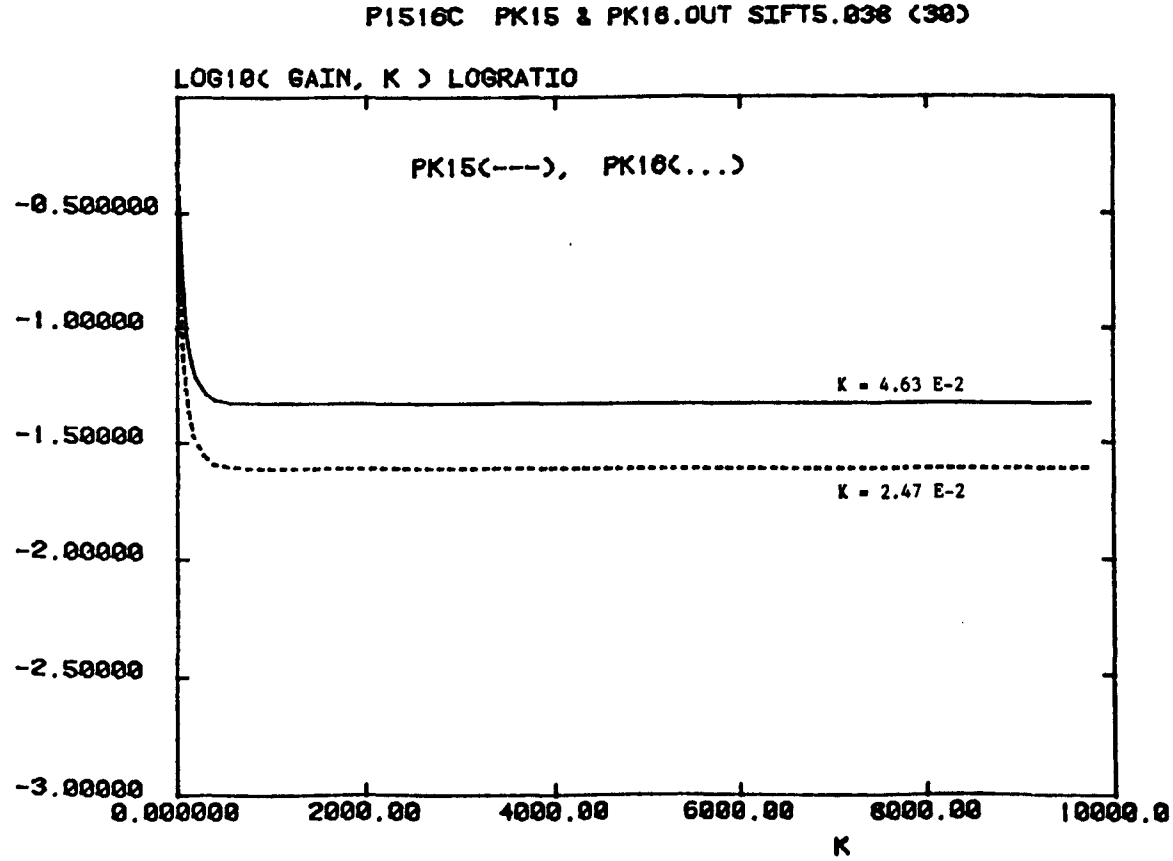
Figure 7.18 (c)



Error Covariance Estimates for SIFT 5·038 (above)

Kalman Gains Used With SIFT 5·038 (below)

Figure 7.18 (d)



Innovation Sequence acfs for SIFT 5·038

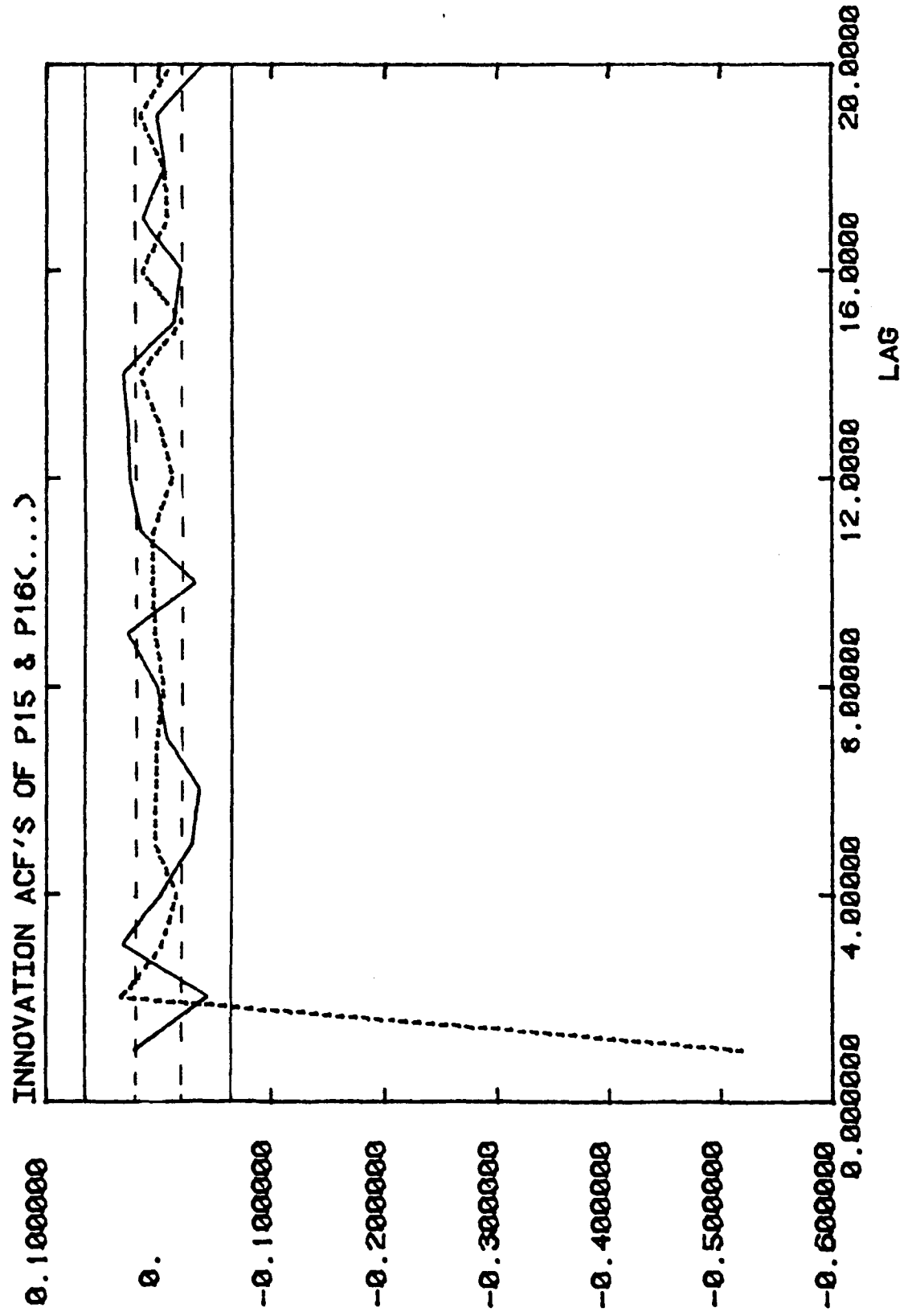
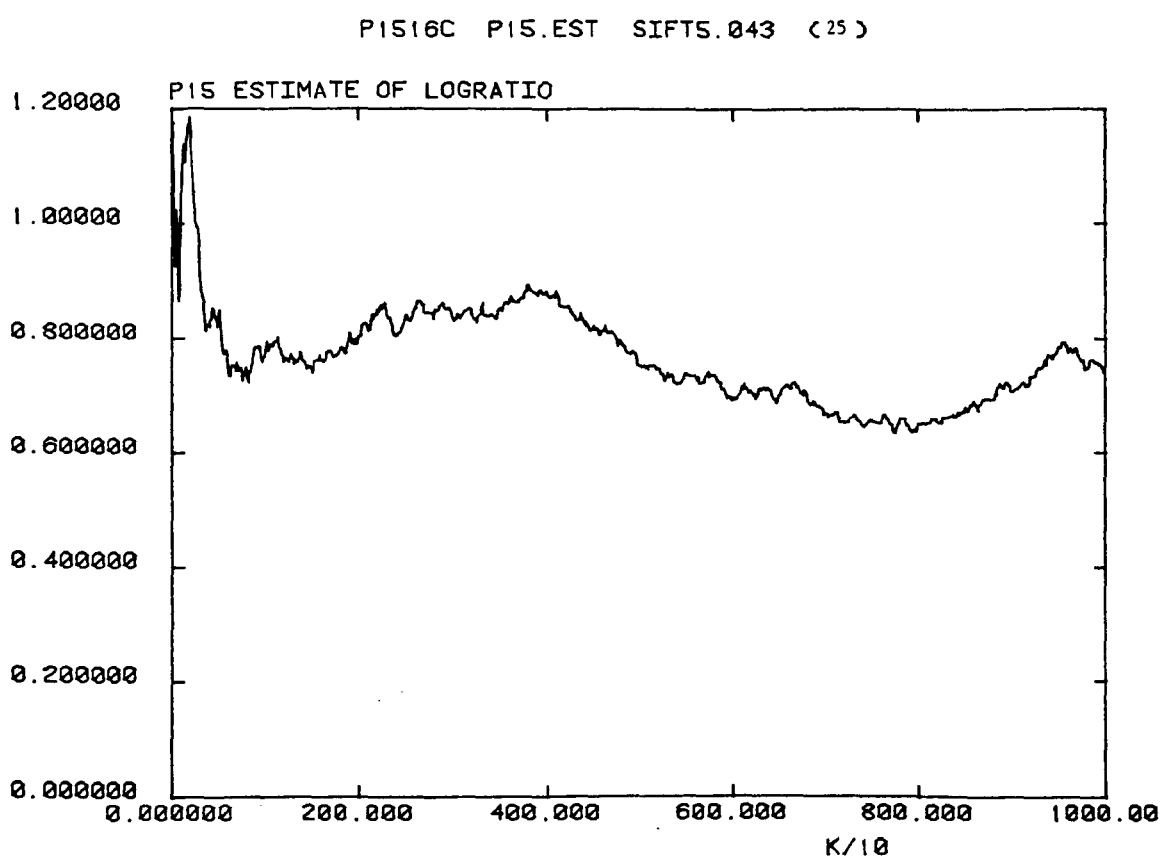


Figure 7.18 (e)

Figure 7.19 (a)



Estimates of the log Ratio for SIFT 5·043 : P15 (above)
: P16 (below)

Figure 7.19 (b)

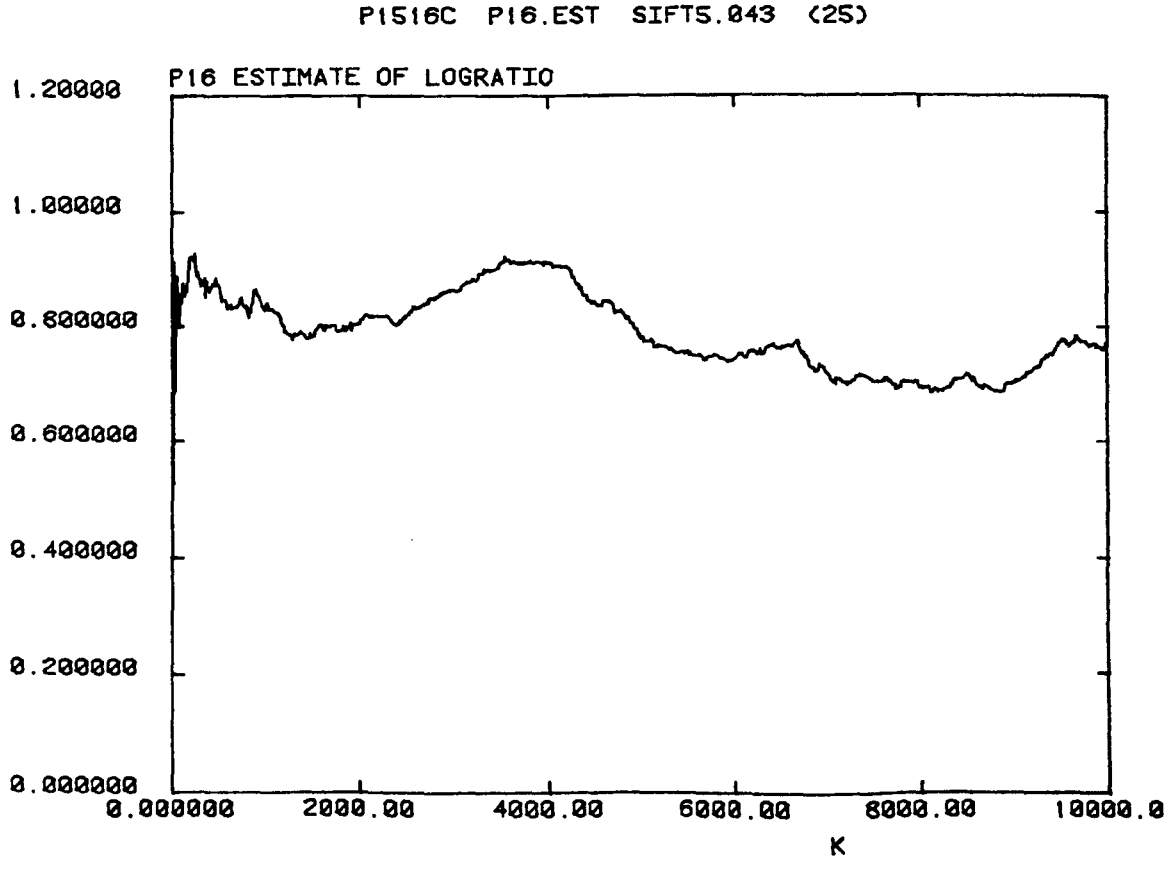
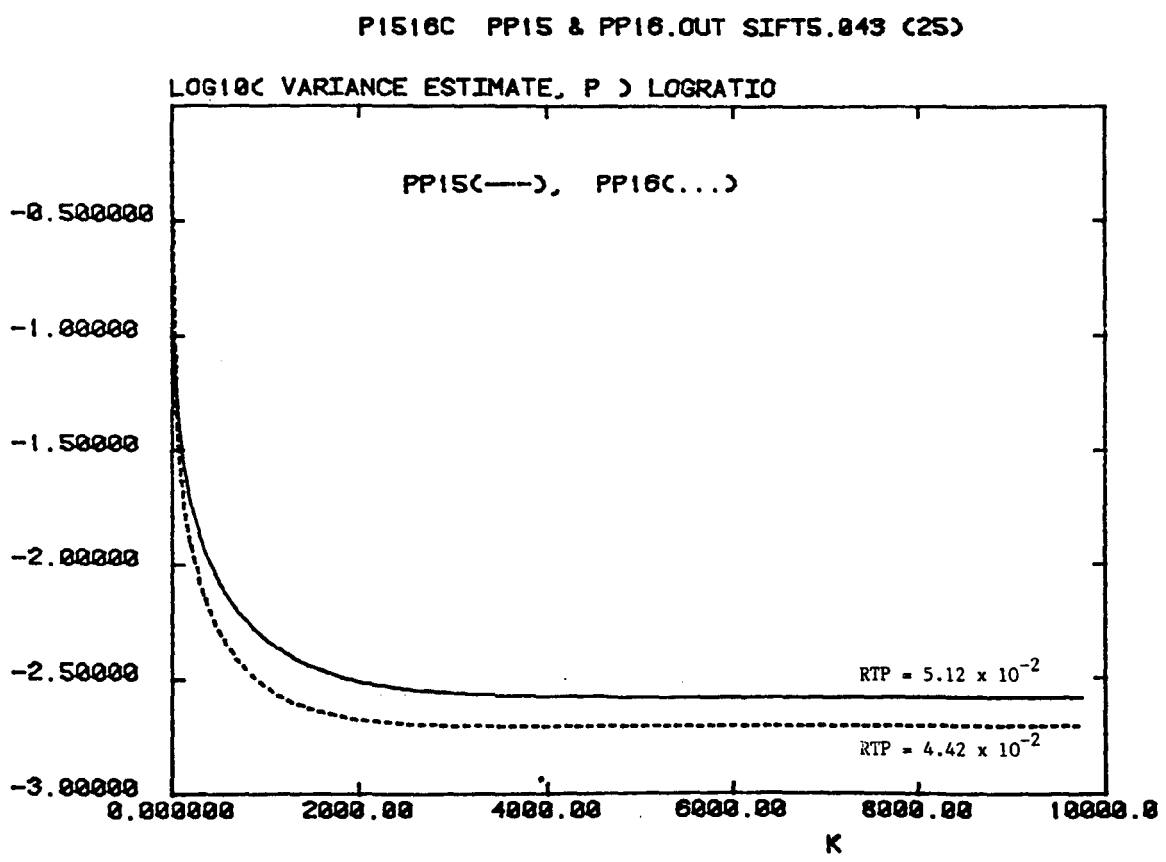


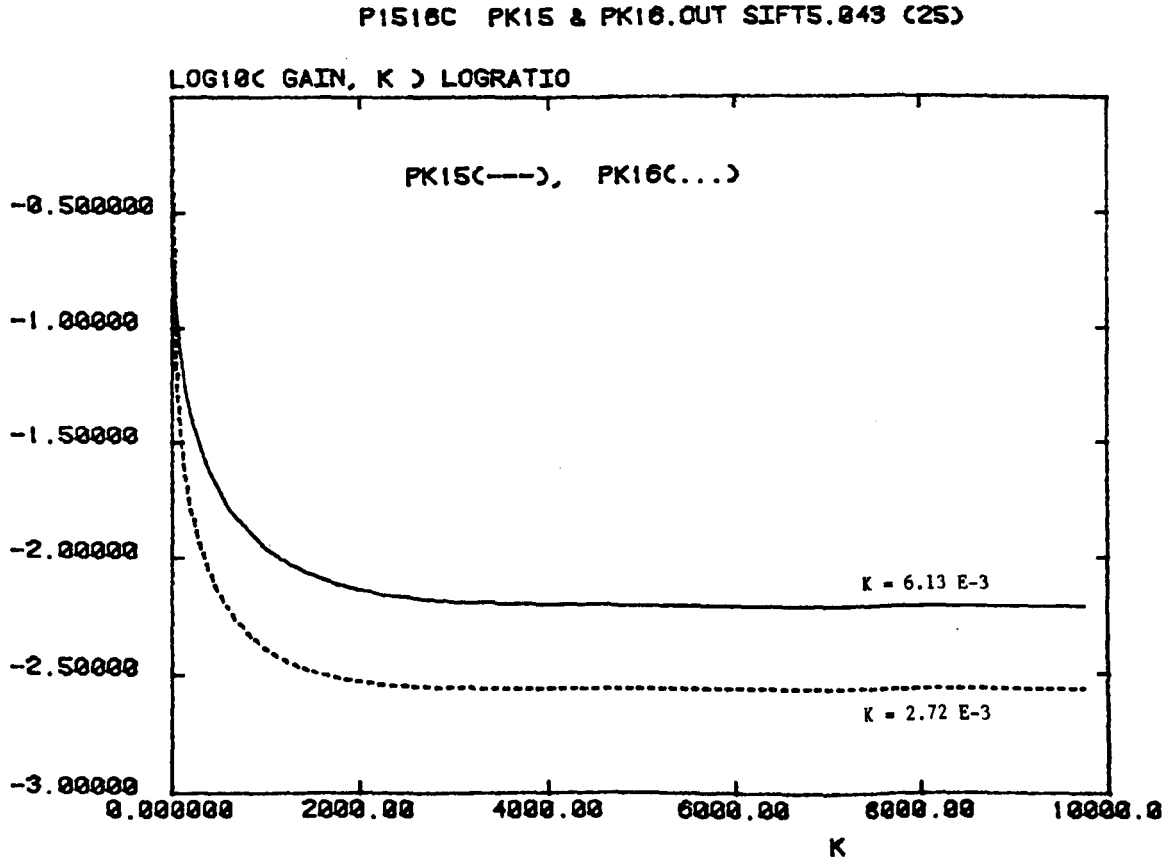
Figure 7.19 (c)



Error Covariance Estimates for SIFT 5·043 (above)

Kalman Gains Used with SIFT 5·043 (below)

Figure 7.19 (d)



Innovation Sequence acfs for SIFT5·043

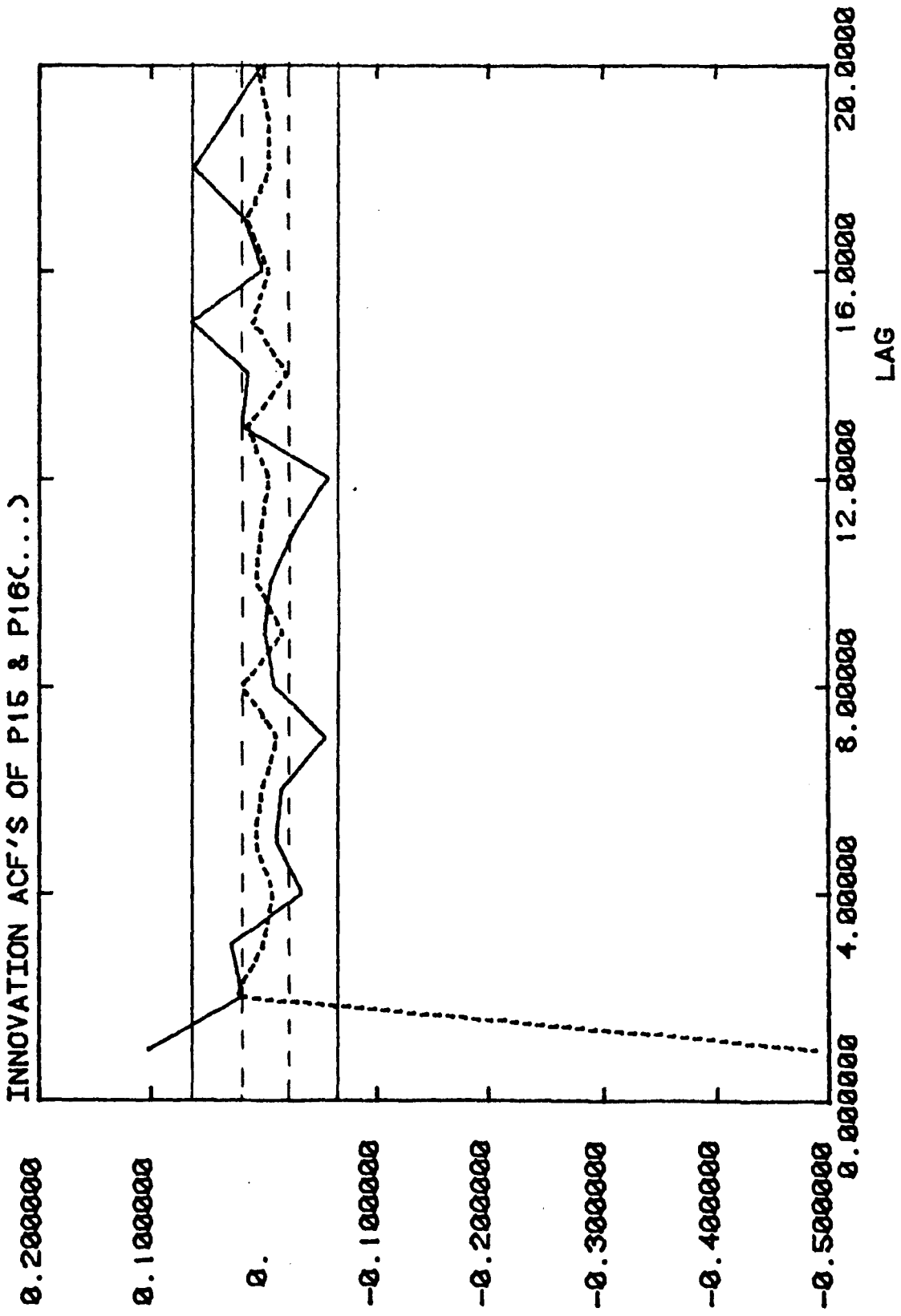
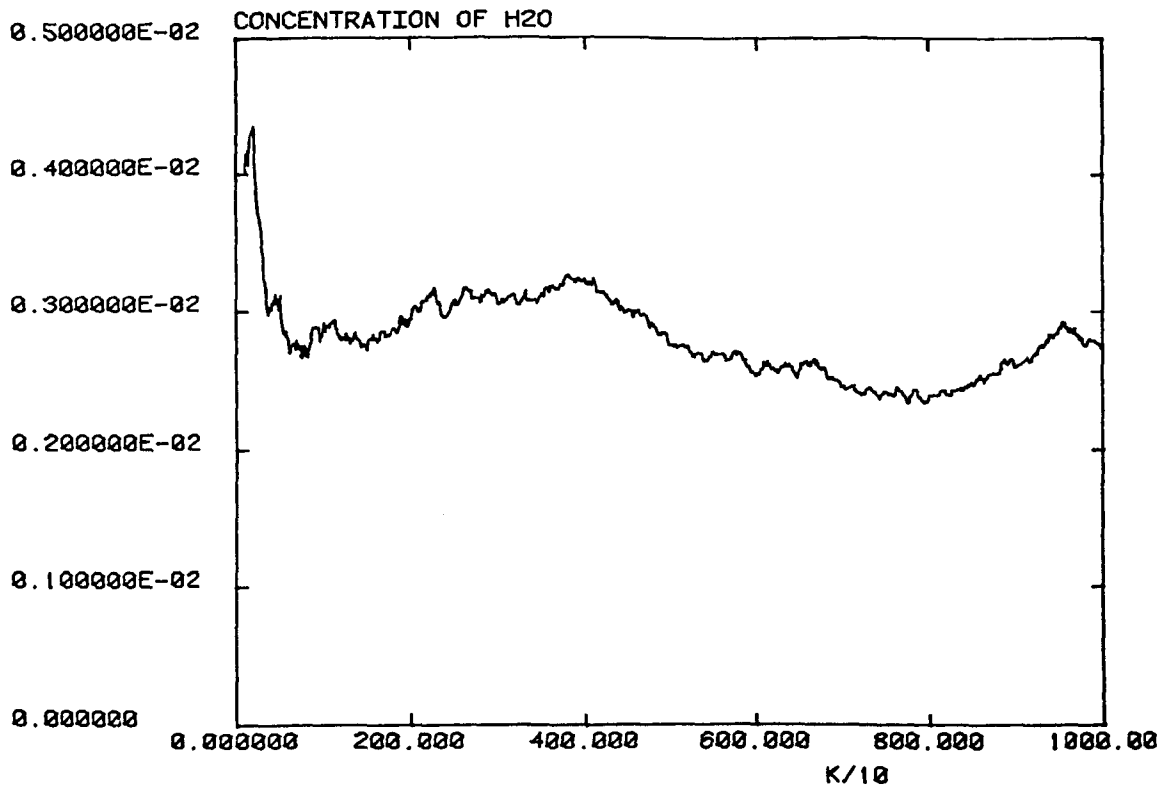


Figure 7.19 (e)

Figure 7.19 (f)

PCONC SIFT5.043 LOGRATIO (25) P15



H_2O Concentration Using Log Ratio Estimates of SIFT5.043:
P.15 (above)
P16 (below)

Figure 7.19 (g)

PCONC SIFT5.043 P16 (25)

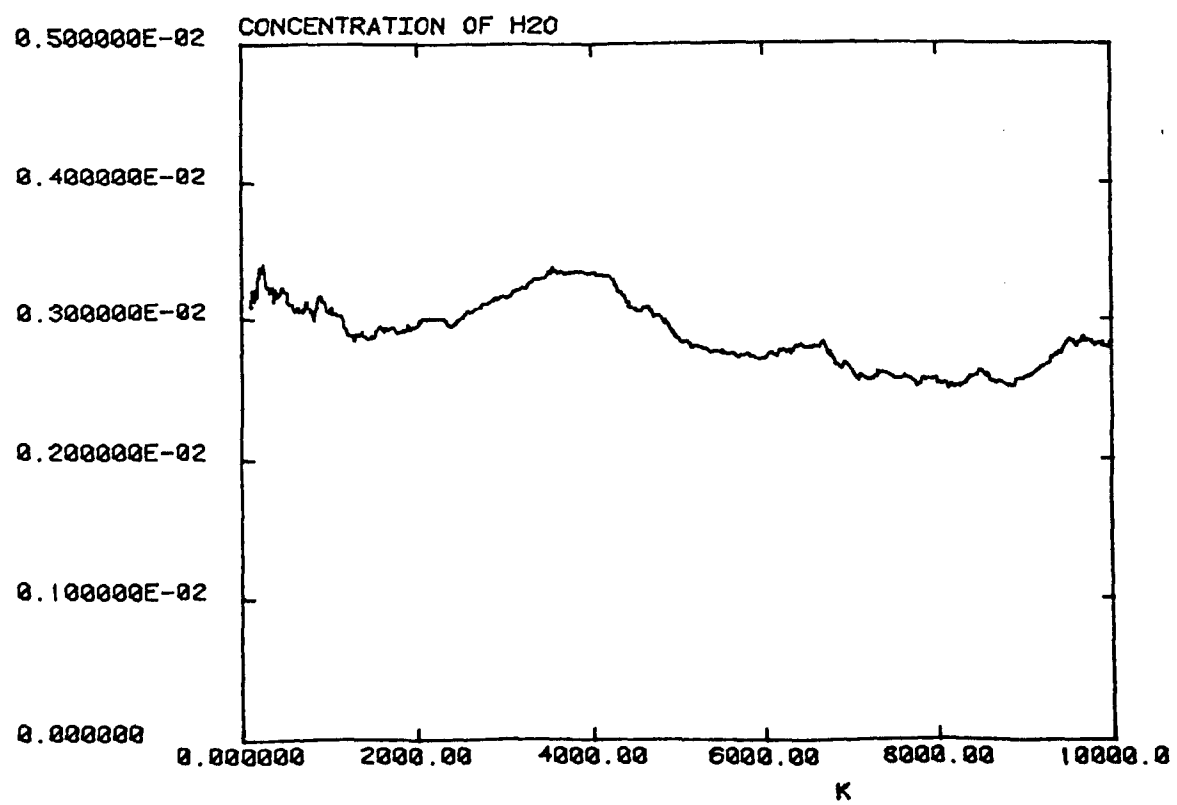
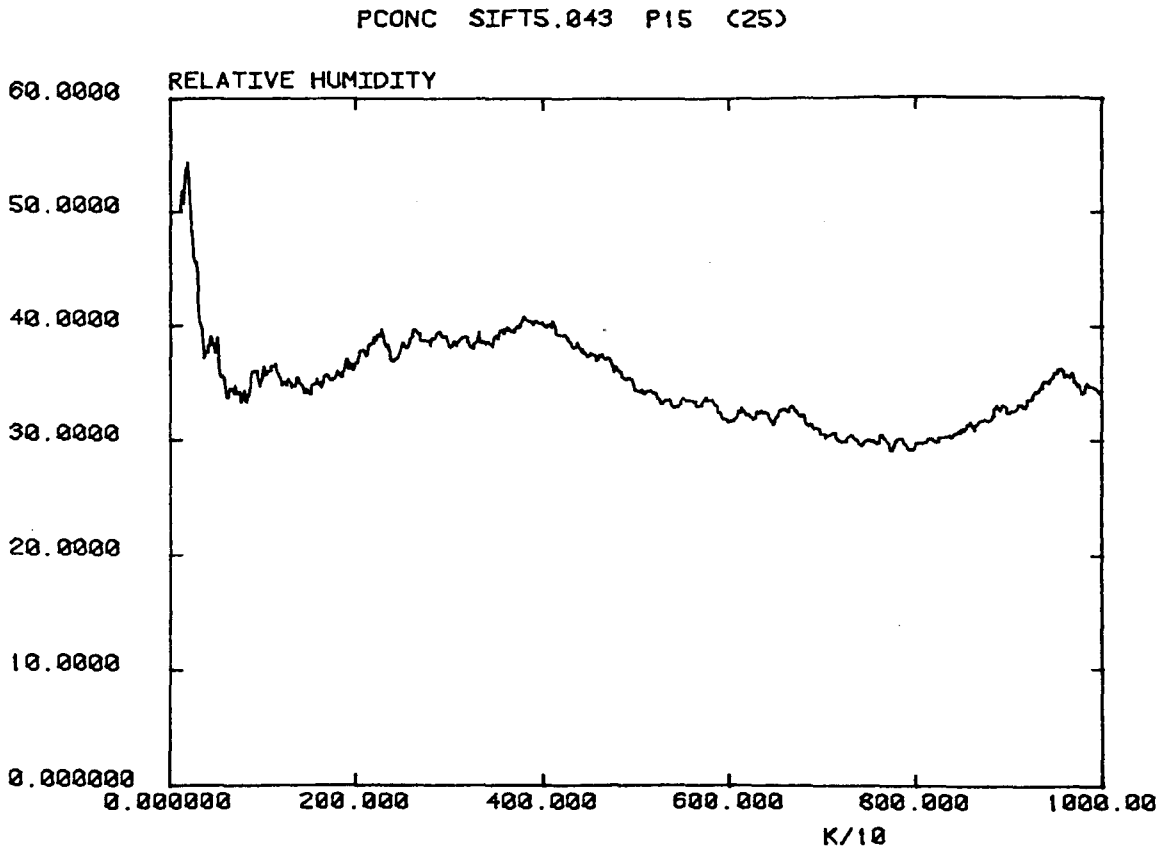


Figure 7.19 (h)



Relative Humidity Using Log Ratio Estimates of SIFT 5.043: P15 (above)
: P16 (below)

Figure 7.19 (i)

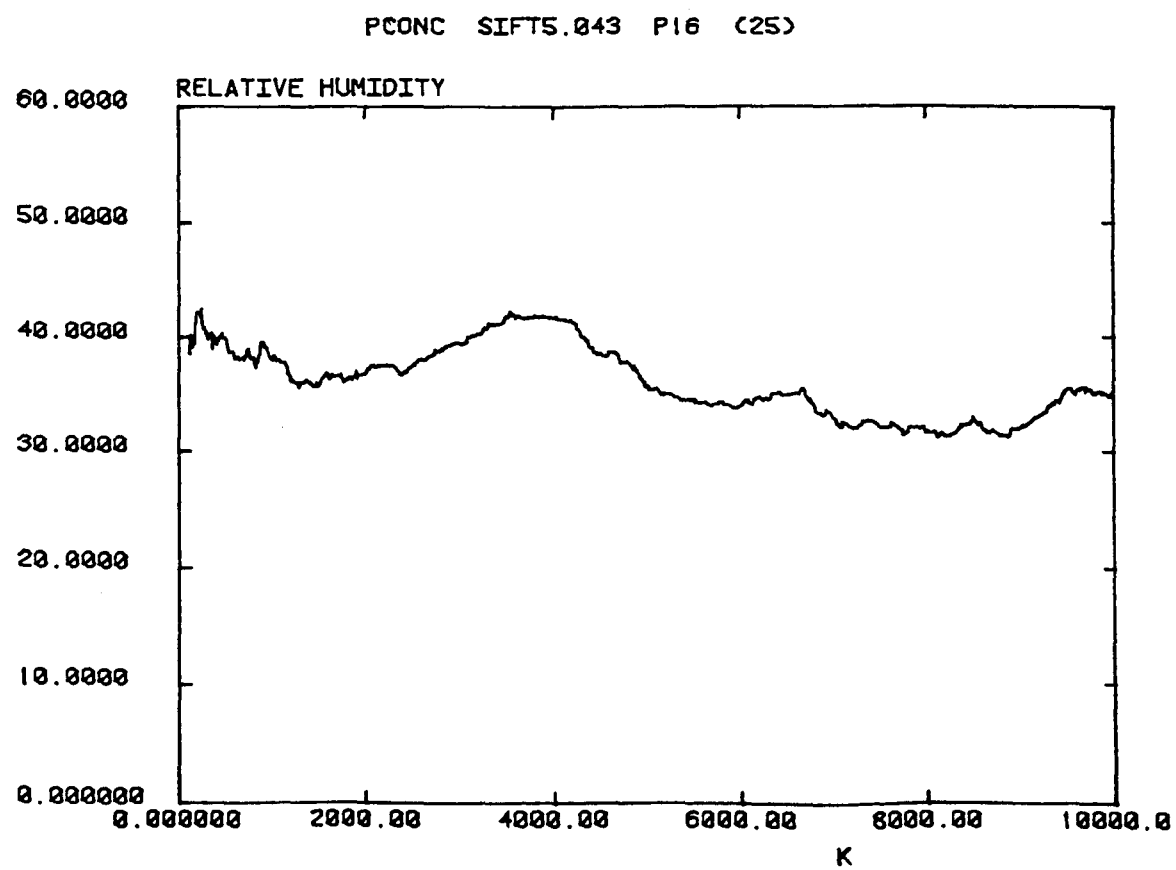
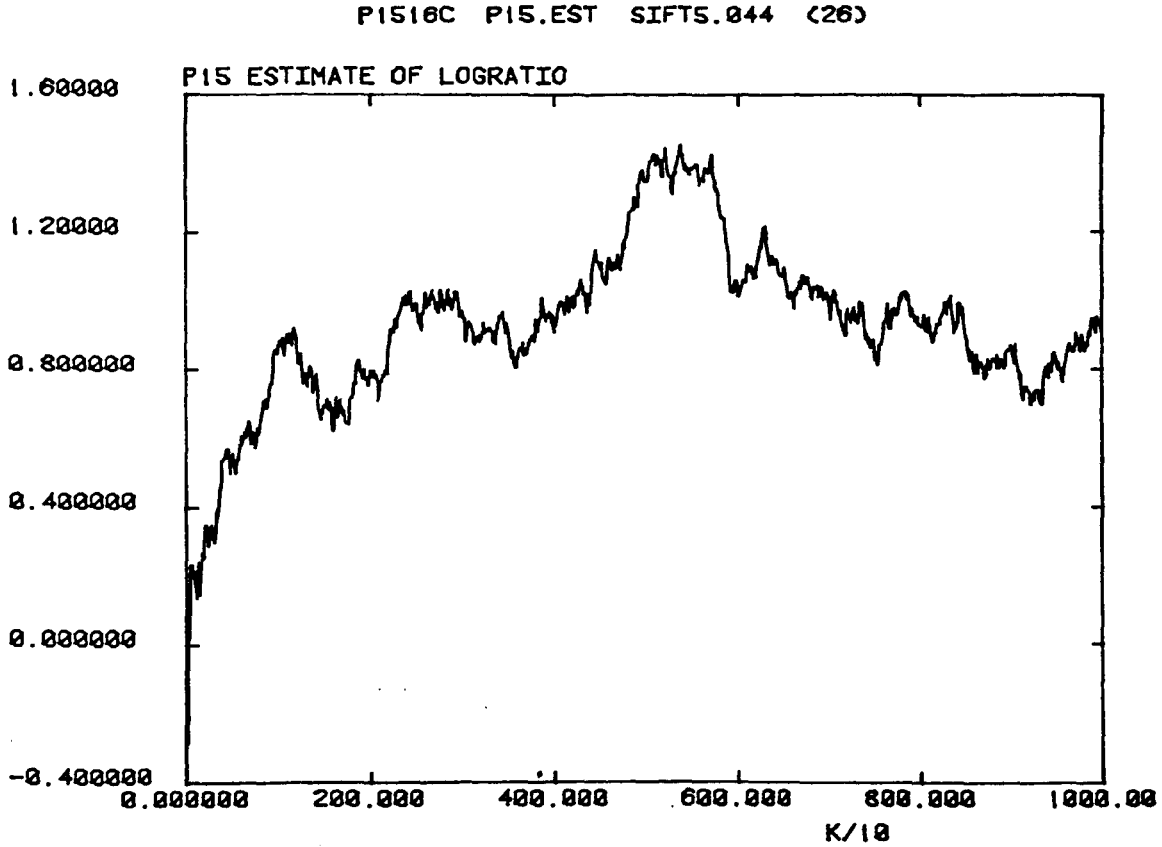


Figure 7.20 (a)



Estimates of the log Ratio for SIFT 5·044 : P15 (above)
: P16 (below)

Figure 7.20 (b)

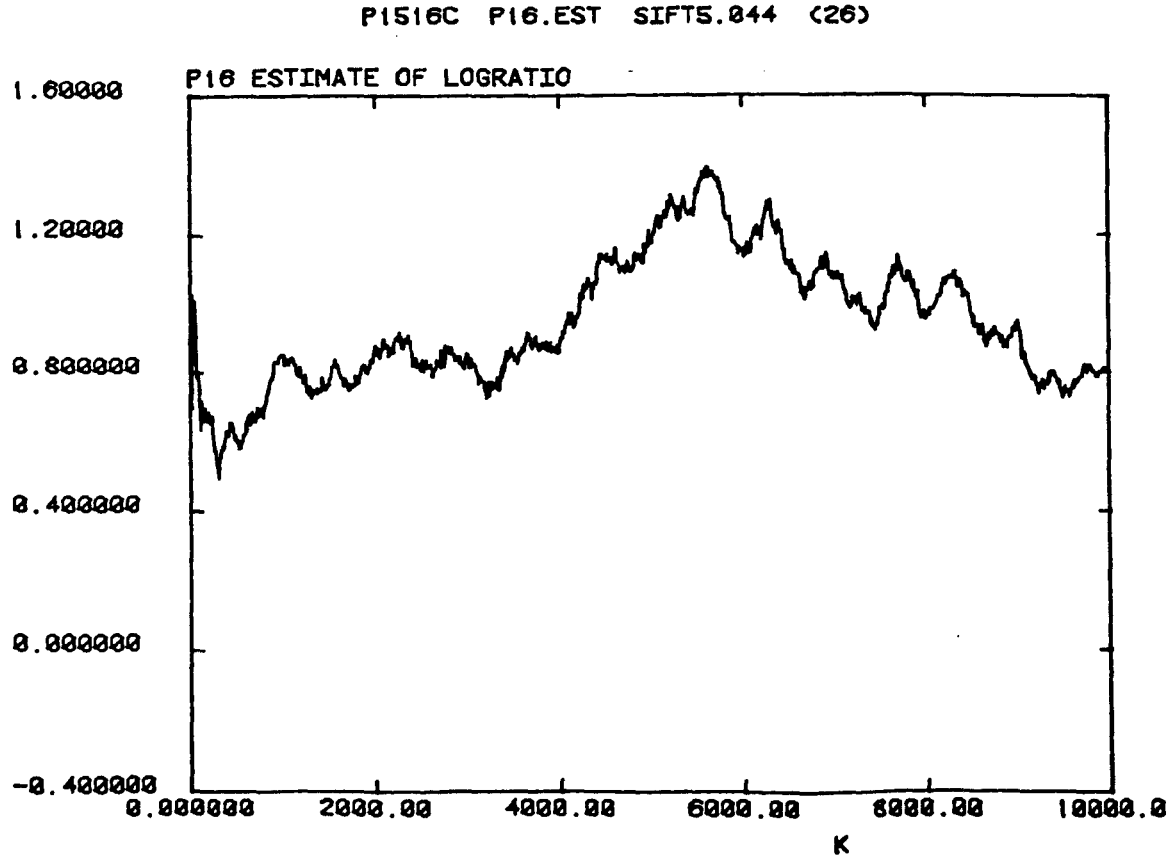
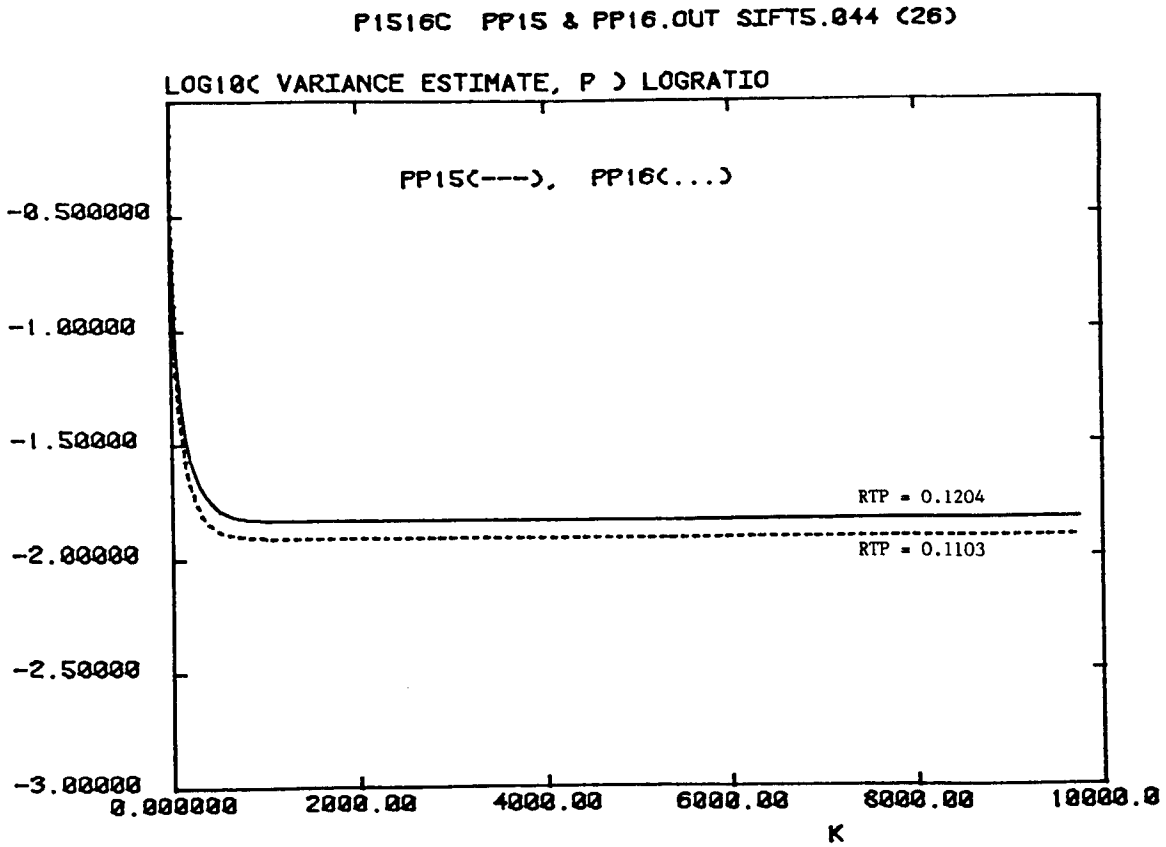


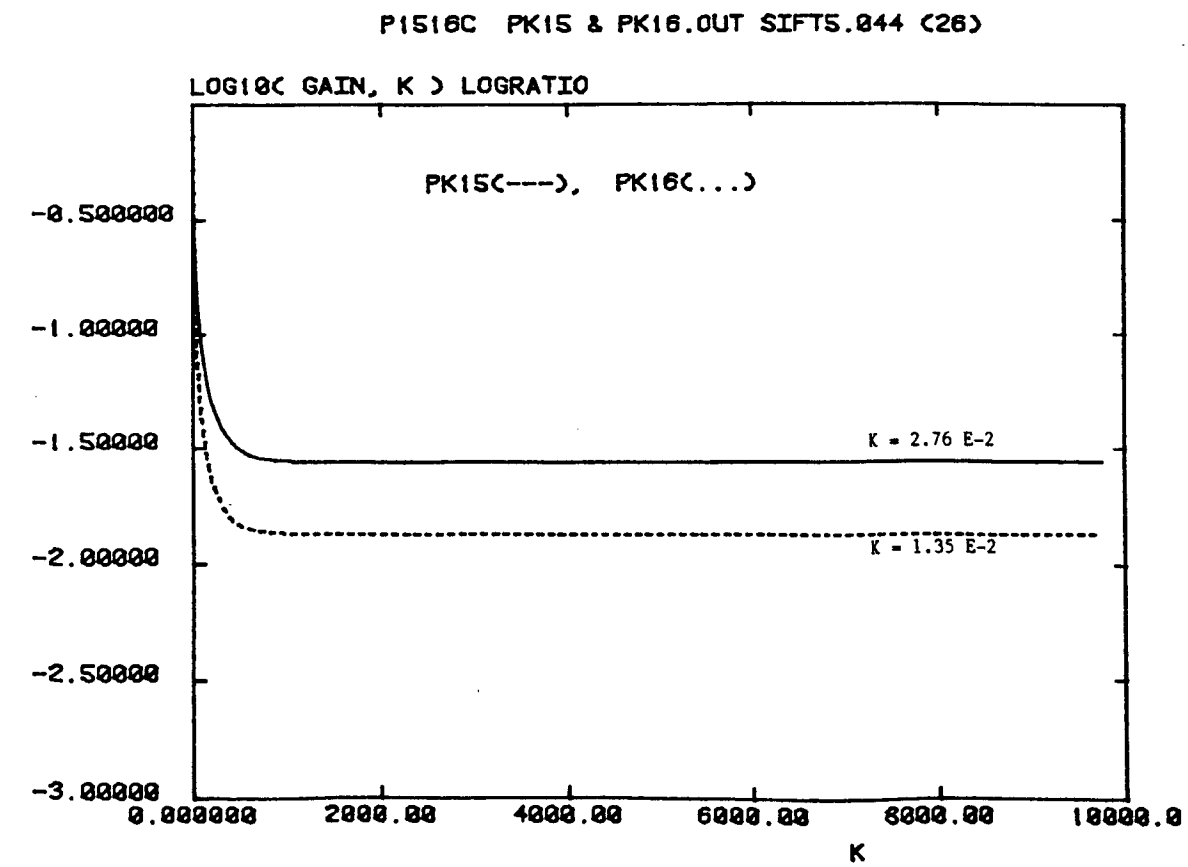
Figure 7.20 (c)



Error Covariance Estimates for SIFT5·044 (above)

Kalman Gains Used With SIFT5·044 (below)

Figure 7.20 (d)



Innovation Sequence acfs for SIFT 5·044

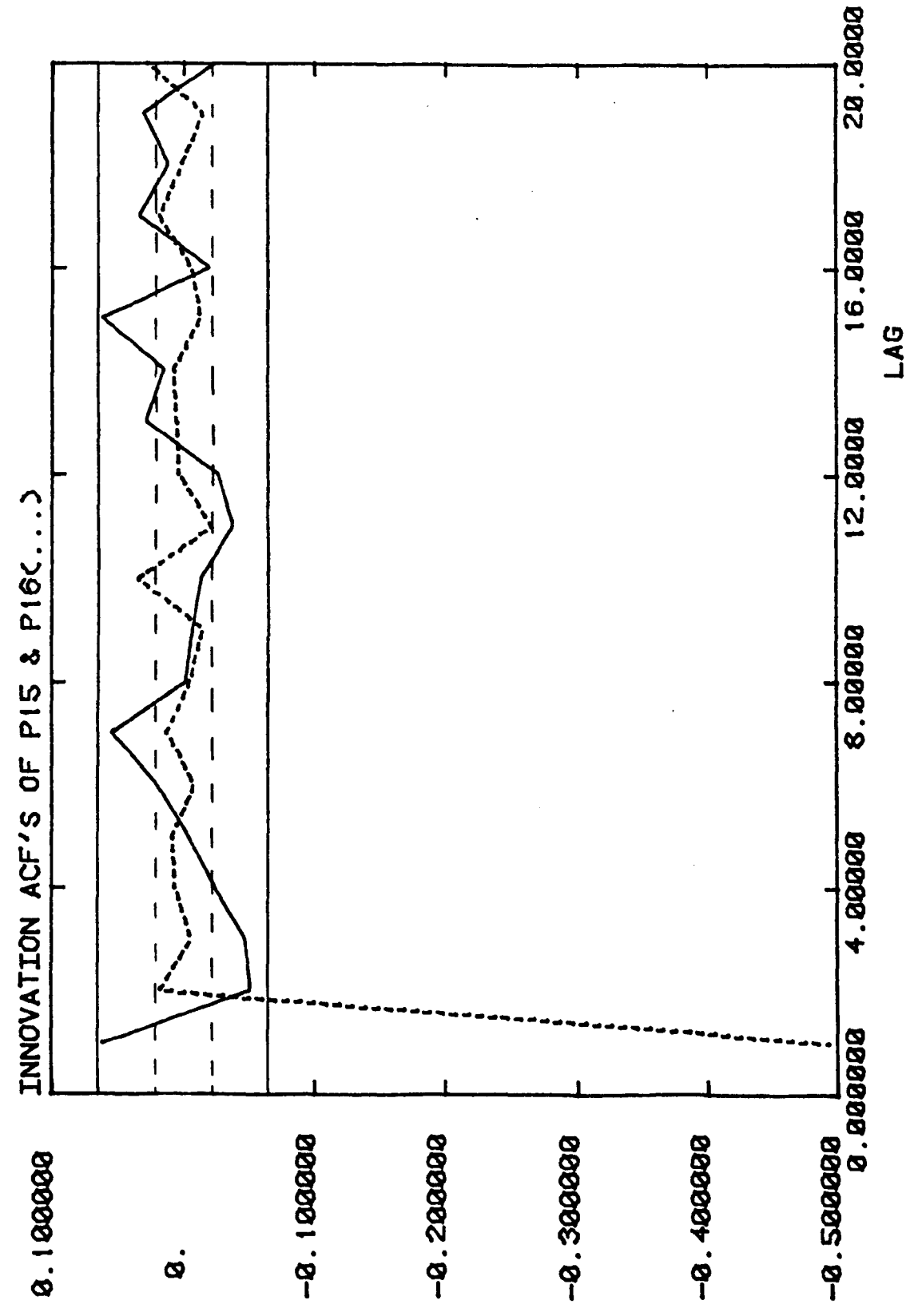
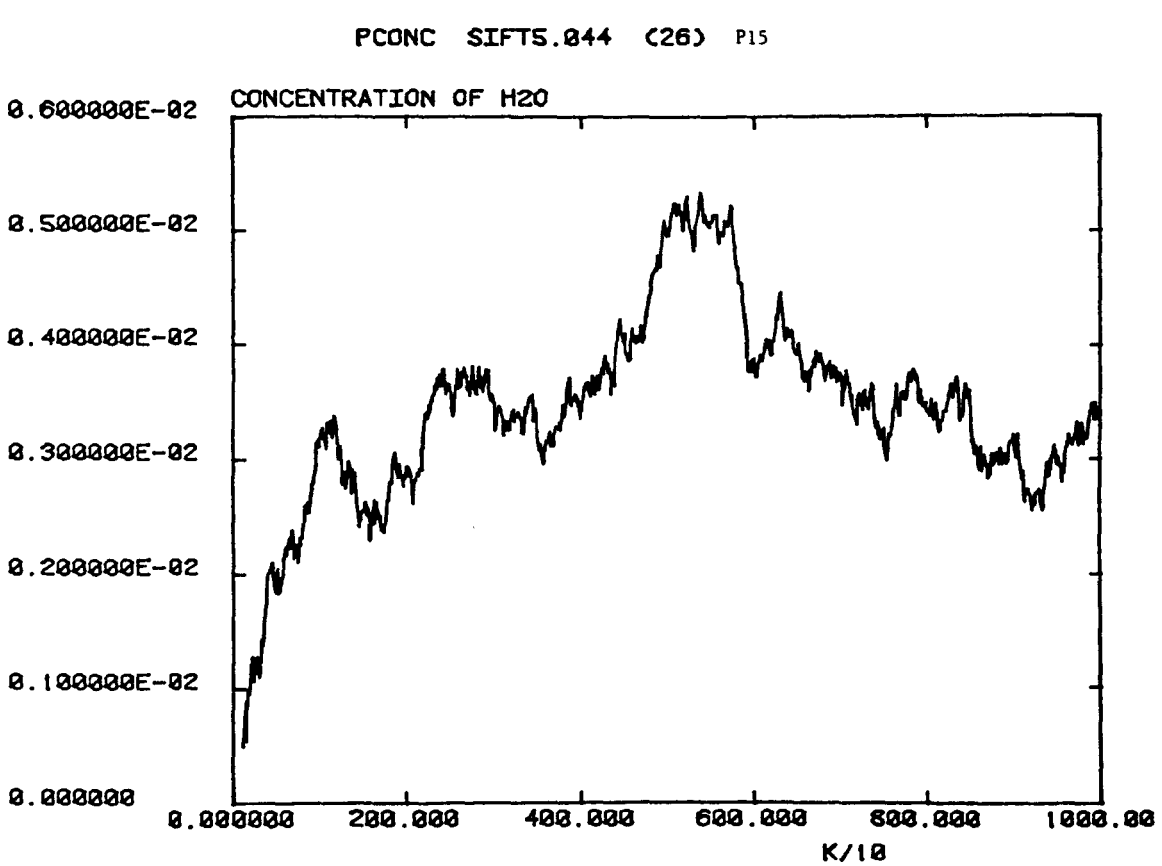


Figure 7.20 (e)

Figure 7.20 (f)



H_2O Concentration Using Log Ratio Estimates of SIFT 5.044:
P15 (above)
P16 (below)

Figure 7.20 (g)

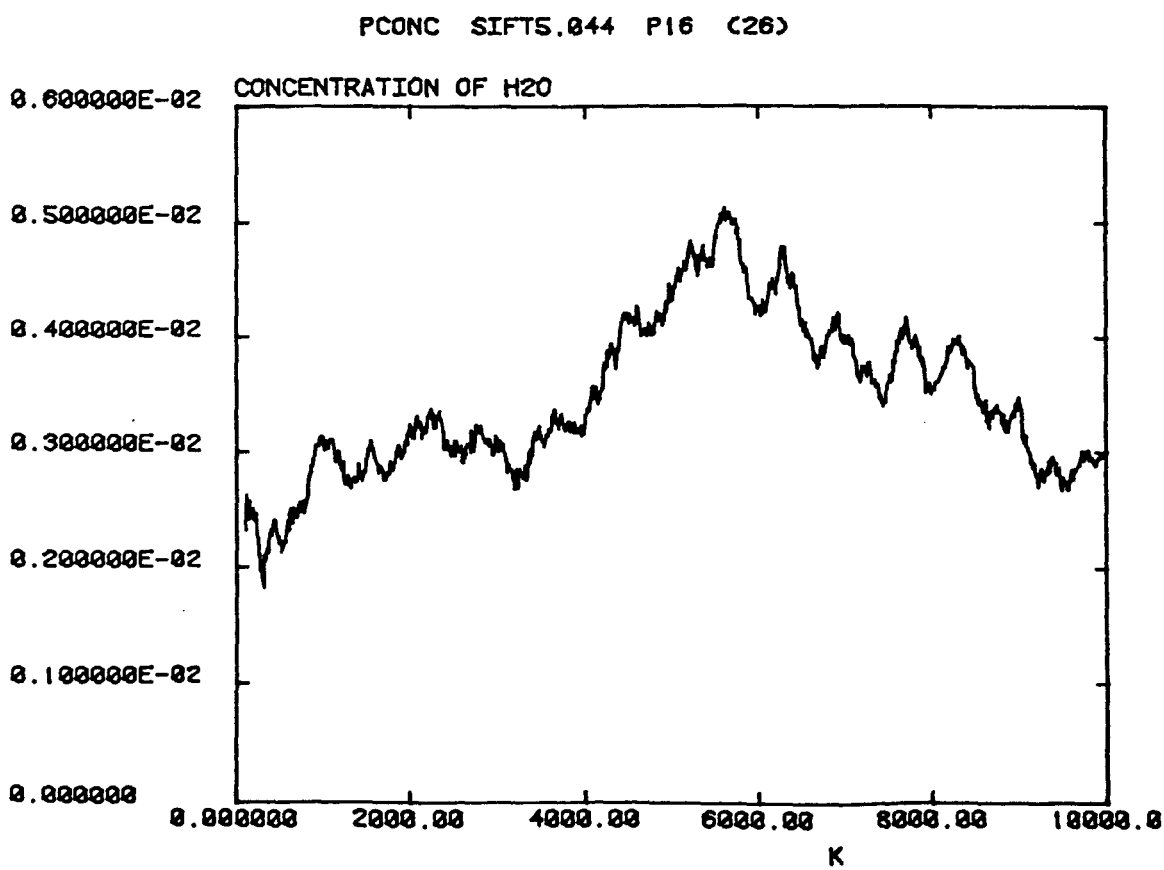
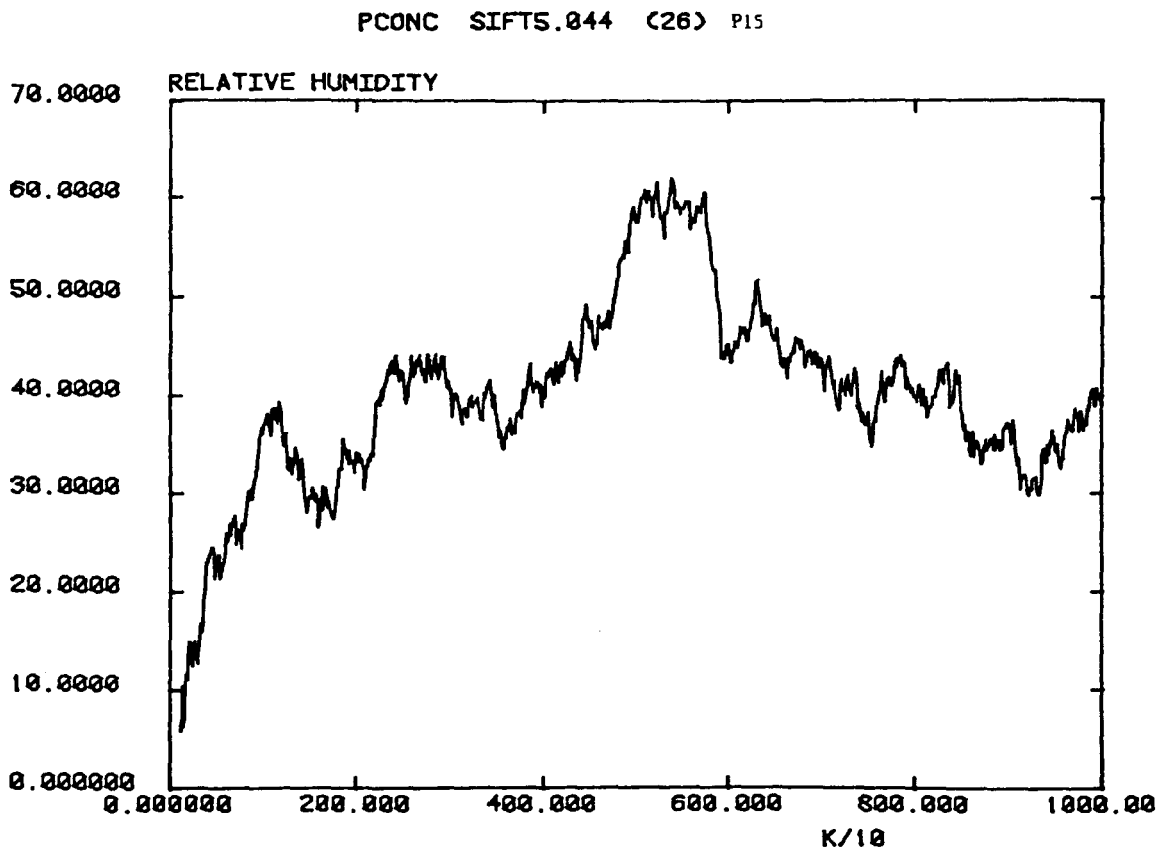


Figure 7.20 (h)



Relative Humidity Using Log Ratio Estimates of SIFT 5.044:
P15 (above)
P16 (below)

Figure 7.20 (i)

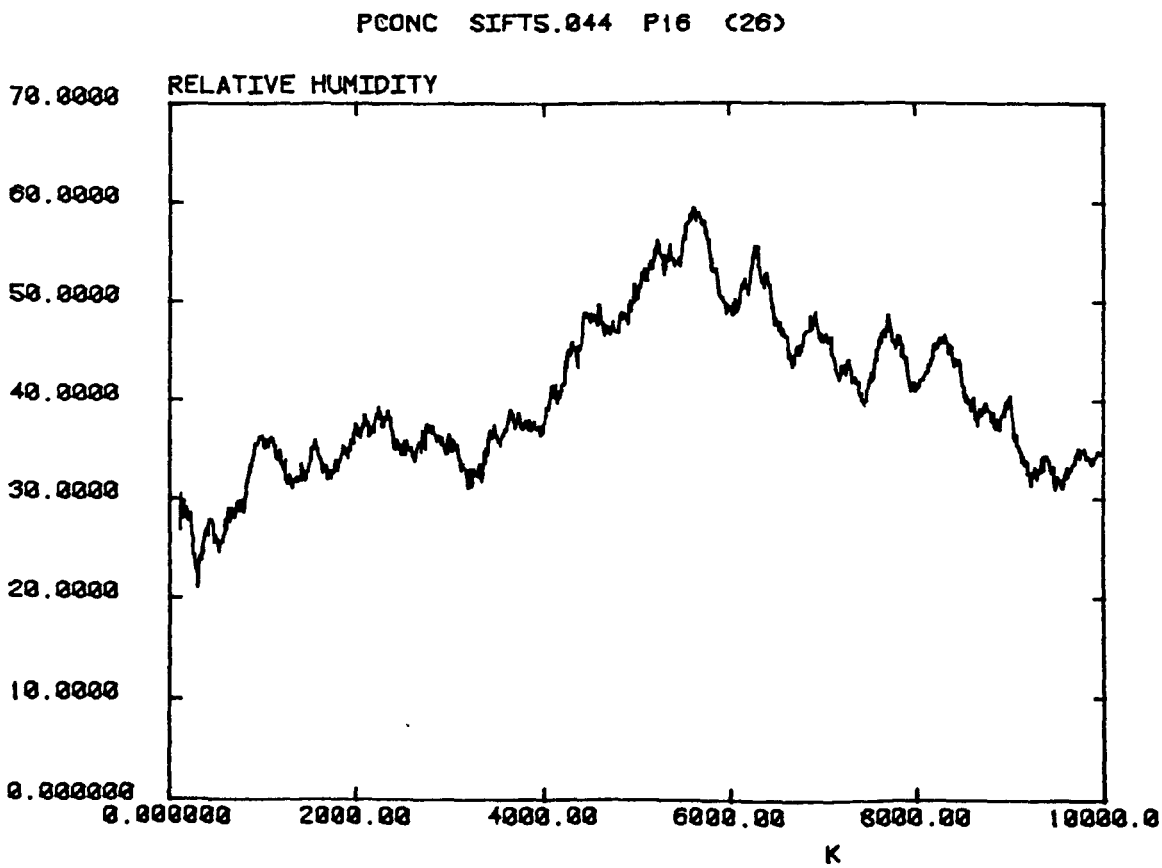
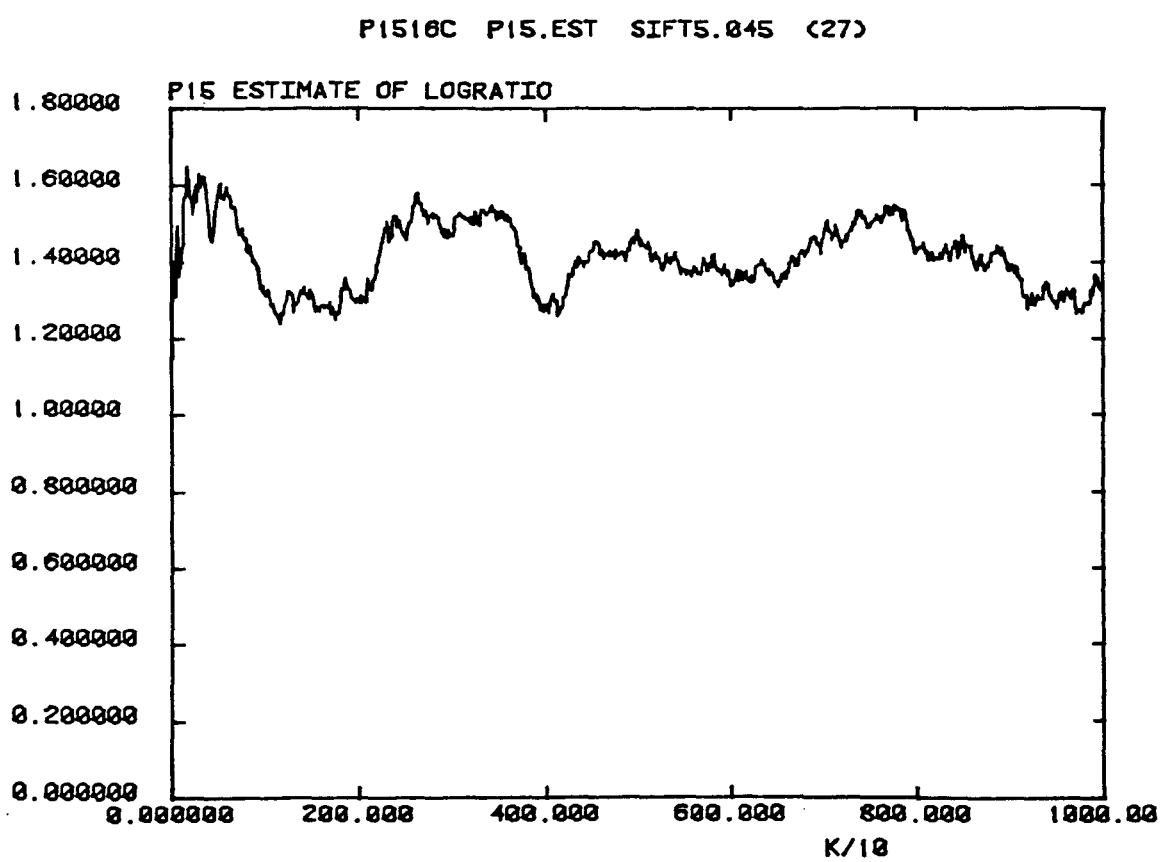


Figure 7.21 (a)



Estimates of the Log Ratio for SIFT 5·045 : P15 (above)
: P16 (below)

Figure 7.21 (b)

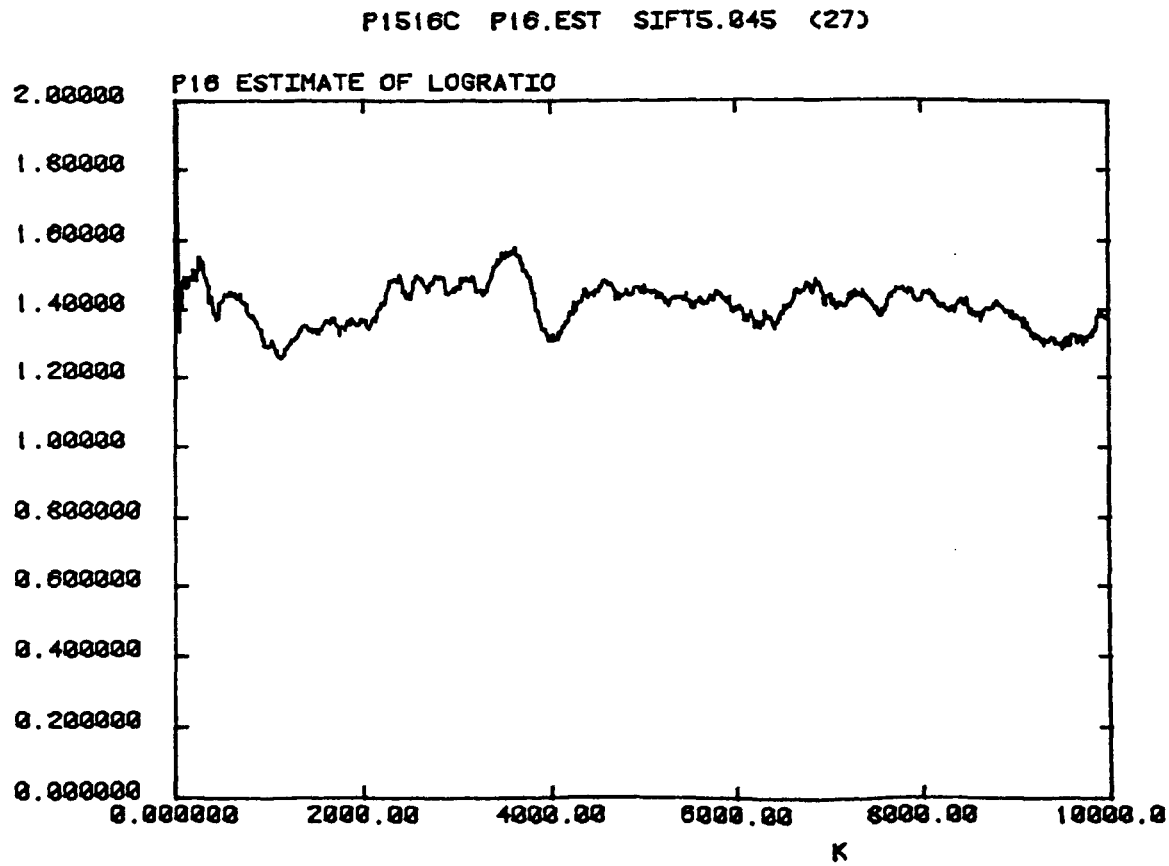
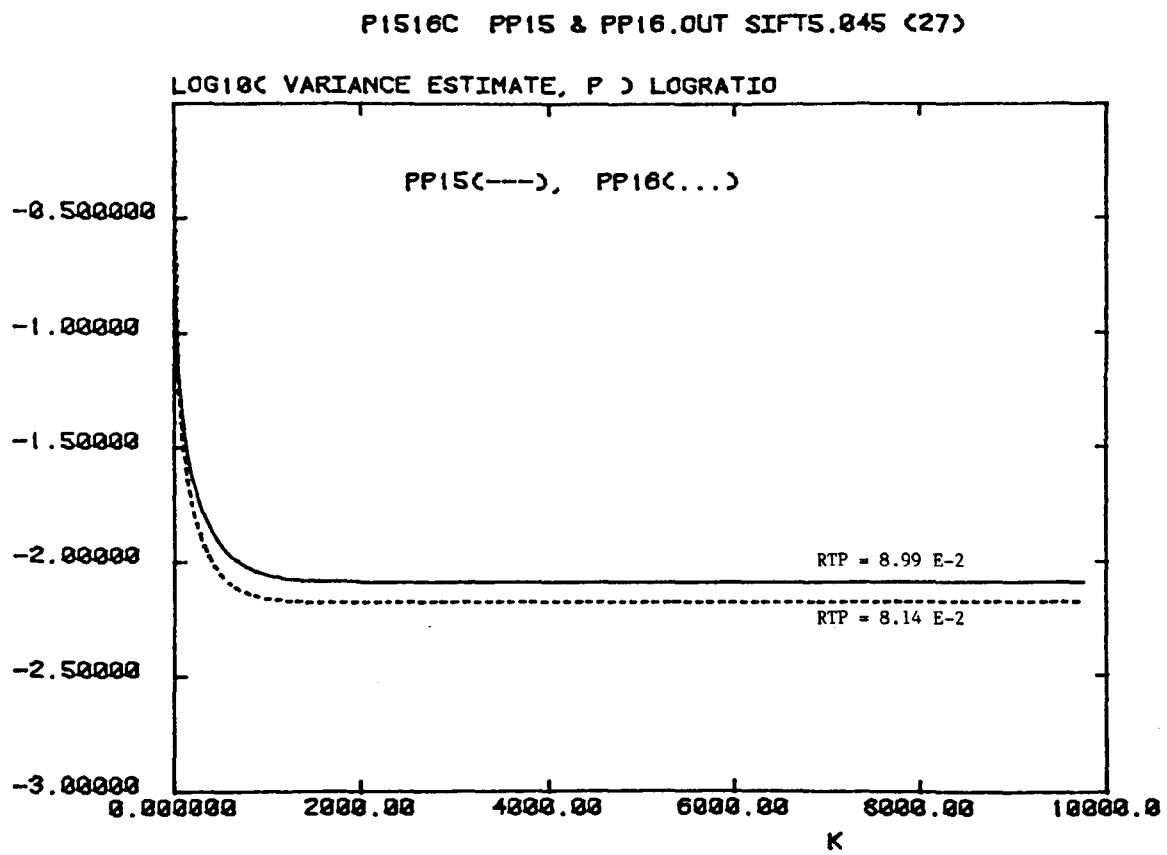


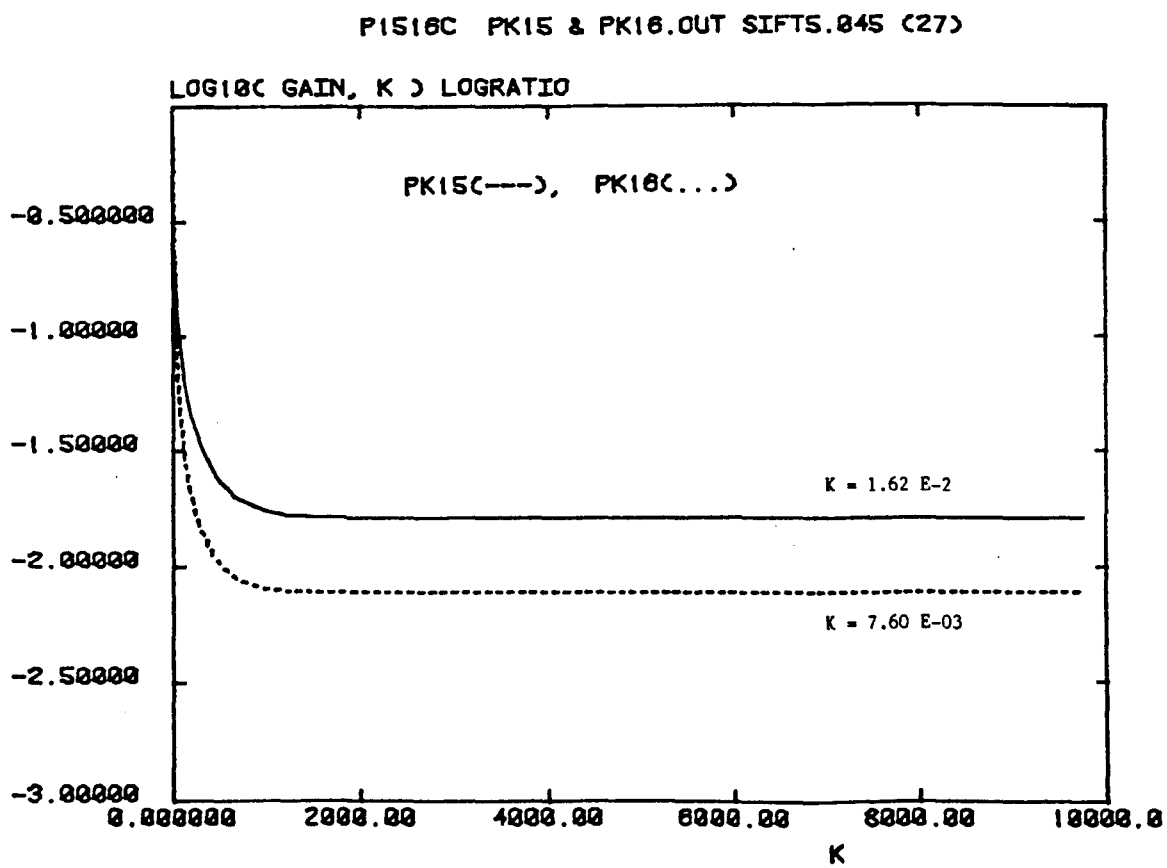
Figure 7.21 (c)



Error Covariance Estimates for SIFT 5.045 (above)

Kalman Gains Used With SIFT 5.045 (below)

Figure 7.21 (d)



Innovation Sequence acfs for SIFT 5.045

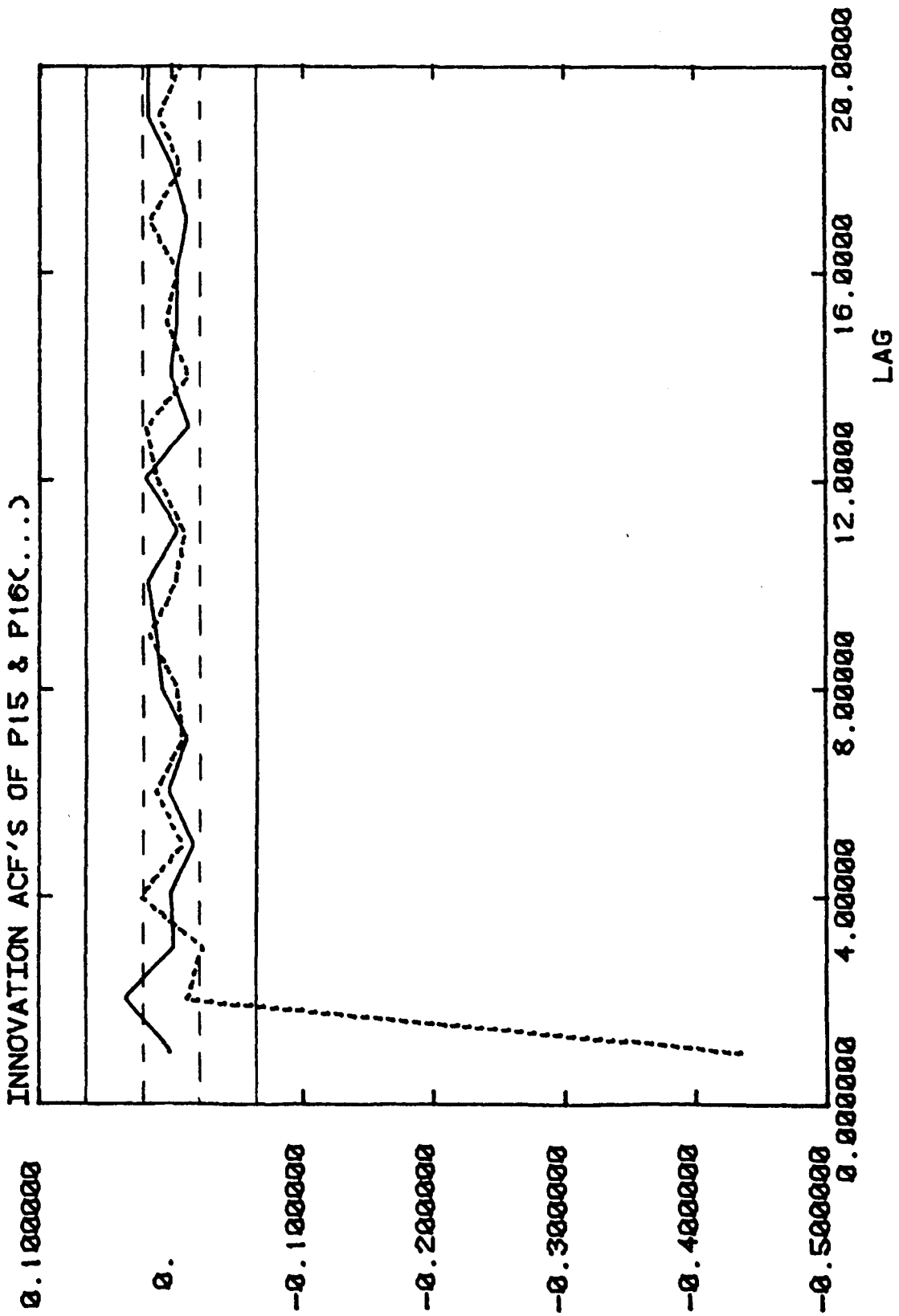
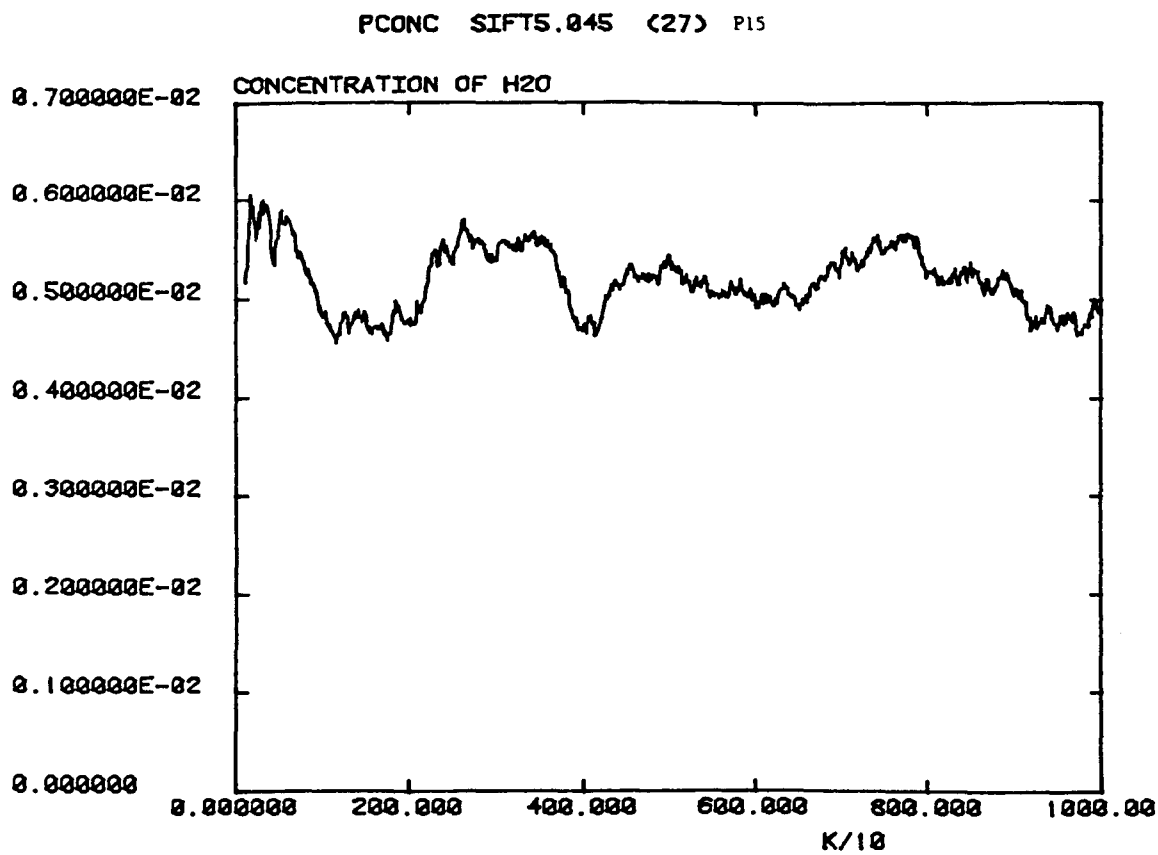


Figure 7.21 (e)

Figure 7.21 (f)



H_2O Concentration Using log Ratio Estimates of SIFT 5·045:

P15 (above)

P16 (below)

Figure 7.21 (g)

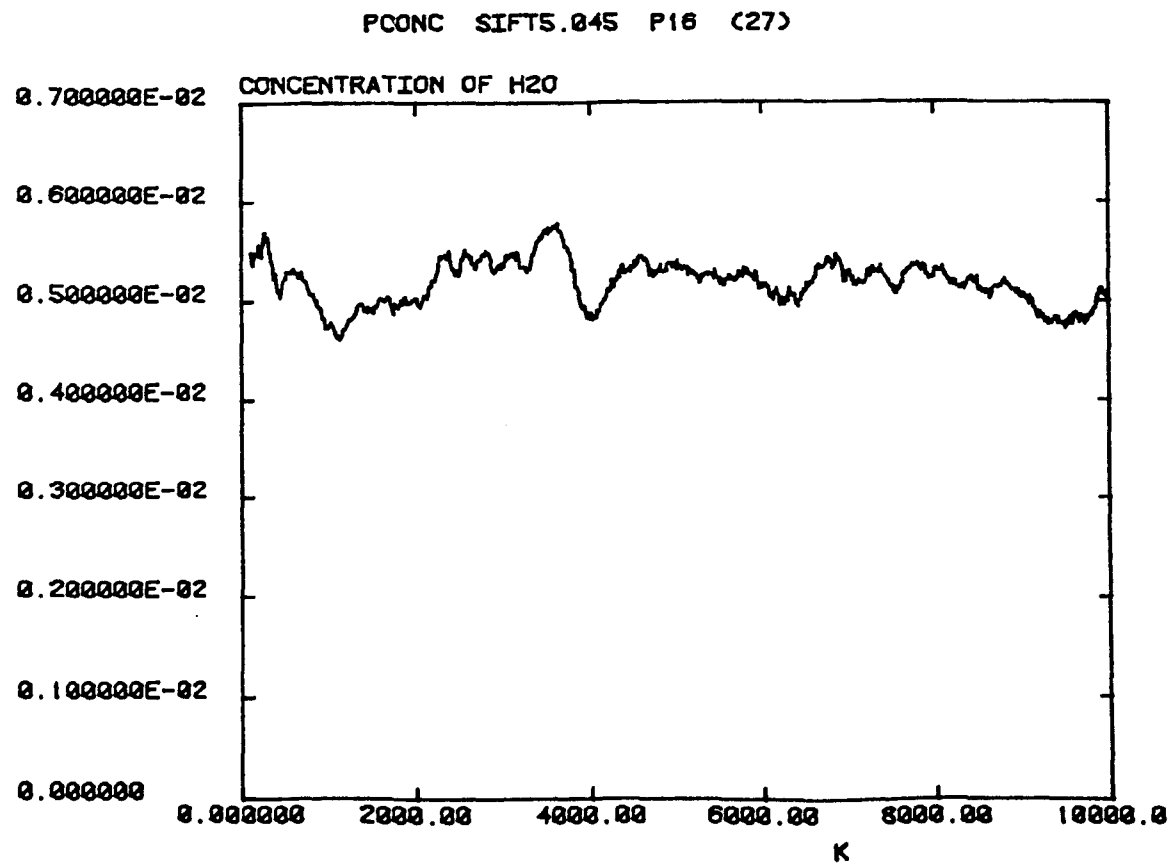
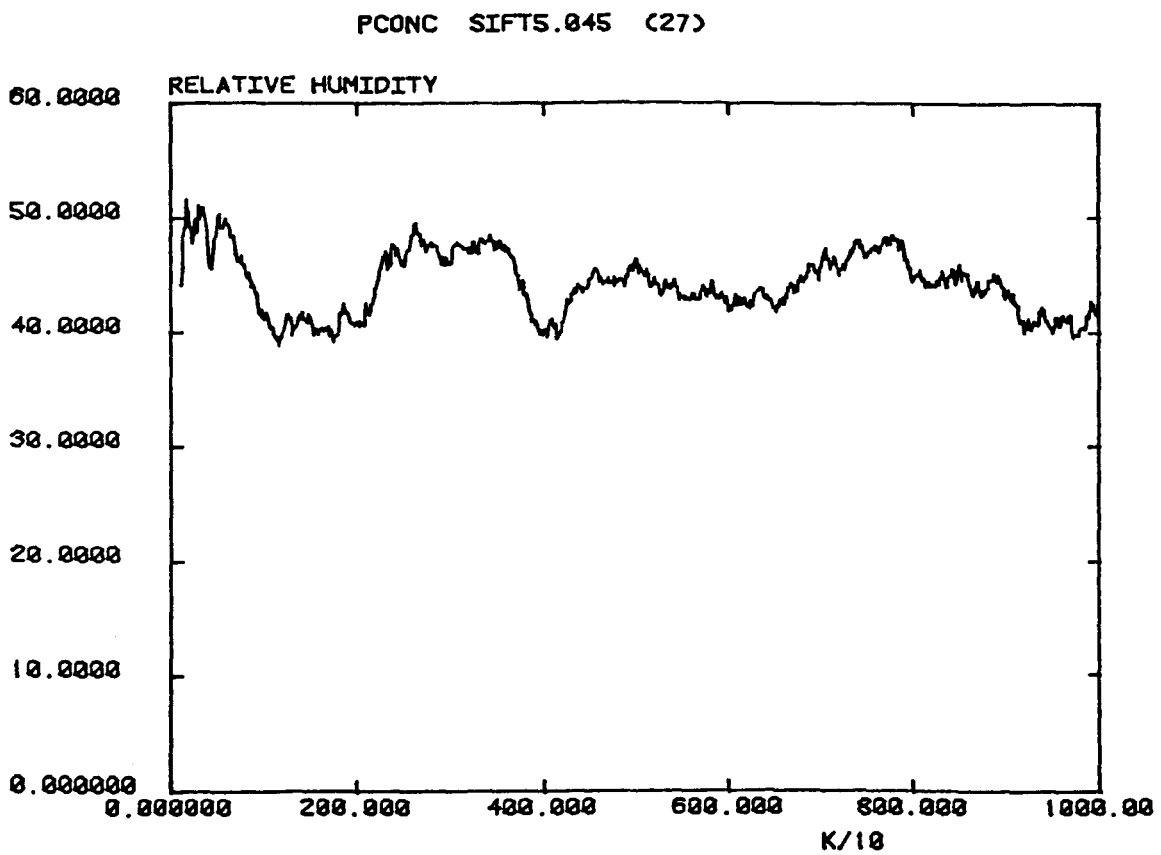


Figure 7.21 (h)



Relative Humidity using Log Ratio Estimates of SIFT 5·045:
P15 (above)
P16 (below)

Figure 7.21 (i)

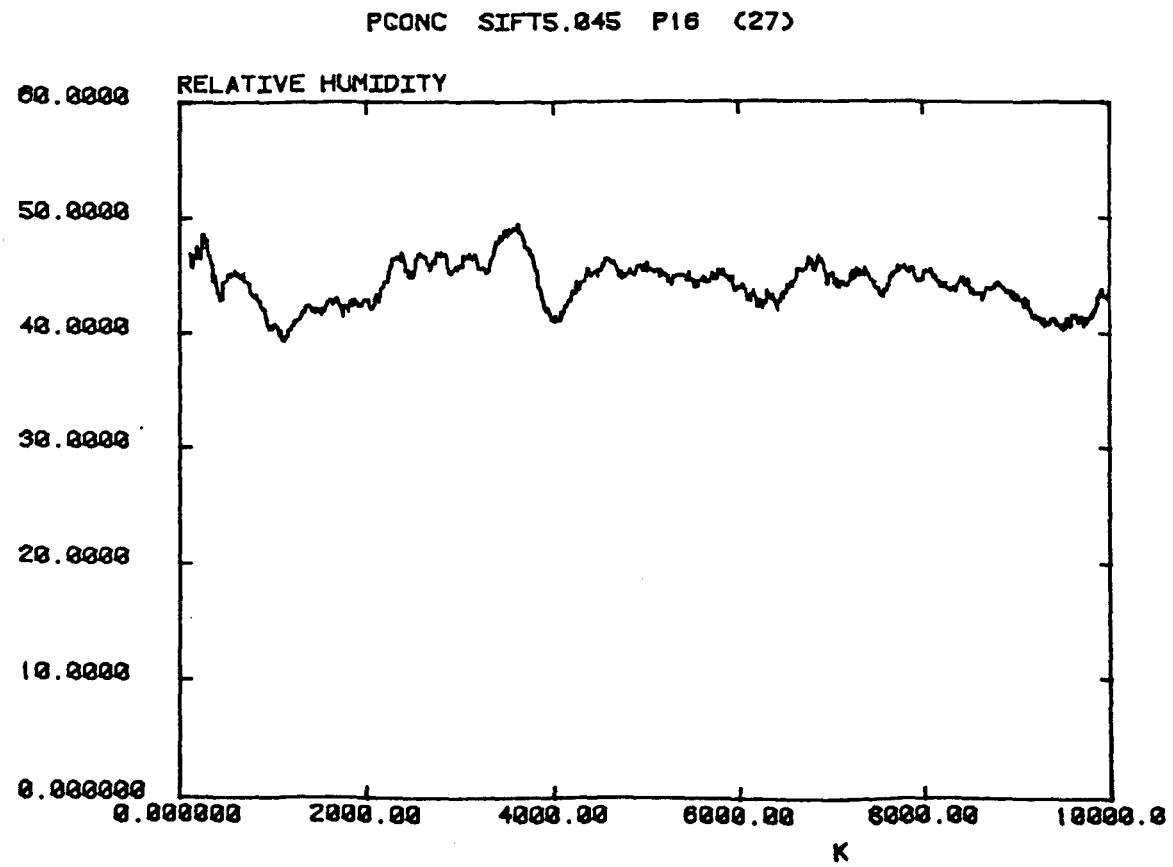
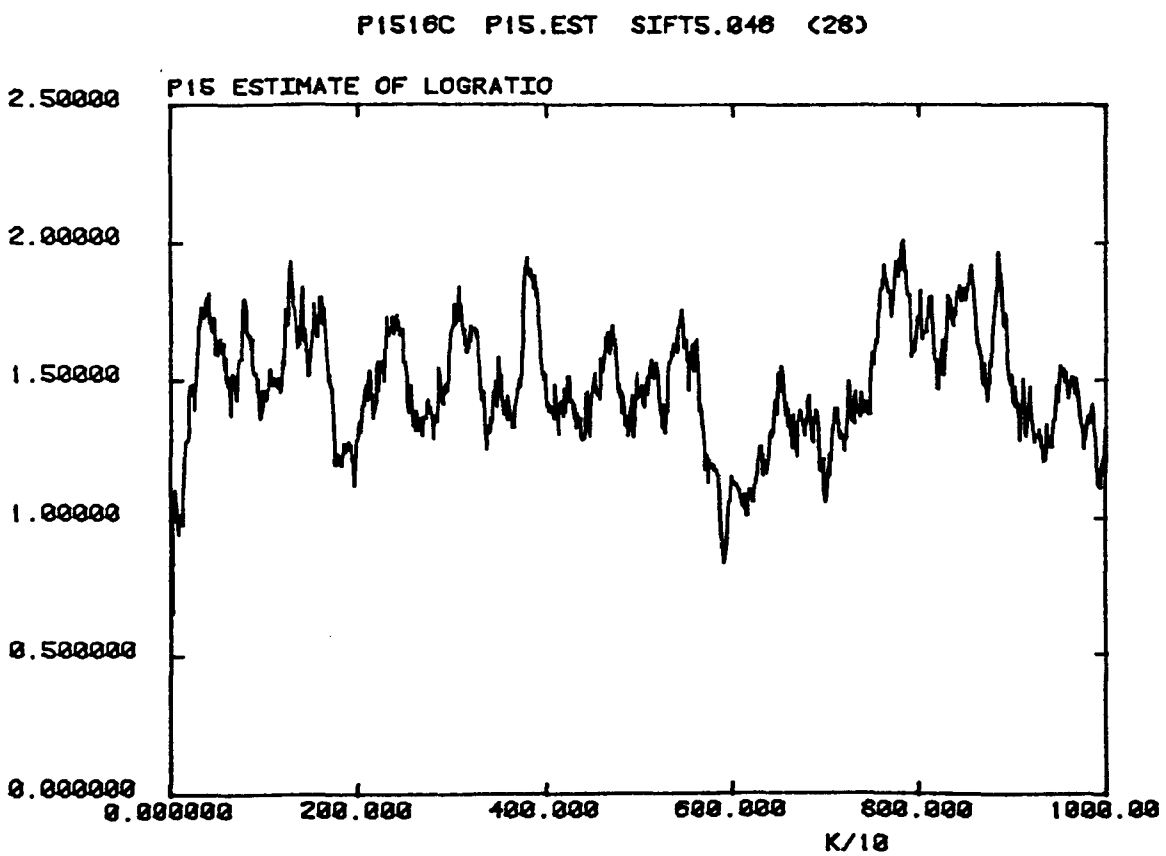


Figure 7.22 (a)



Estimates of the Log Ratio for SIFT 5·046 : P15 (above)

: P16 (below)

Figure 7.22 (b)

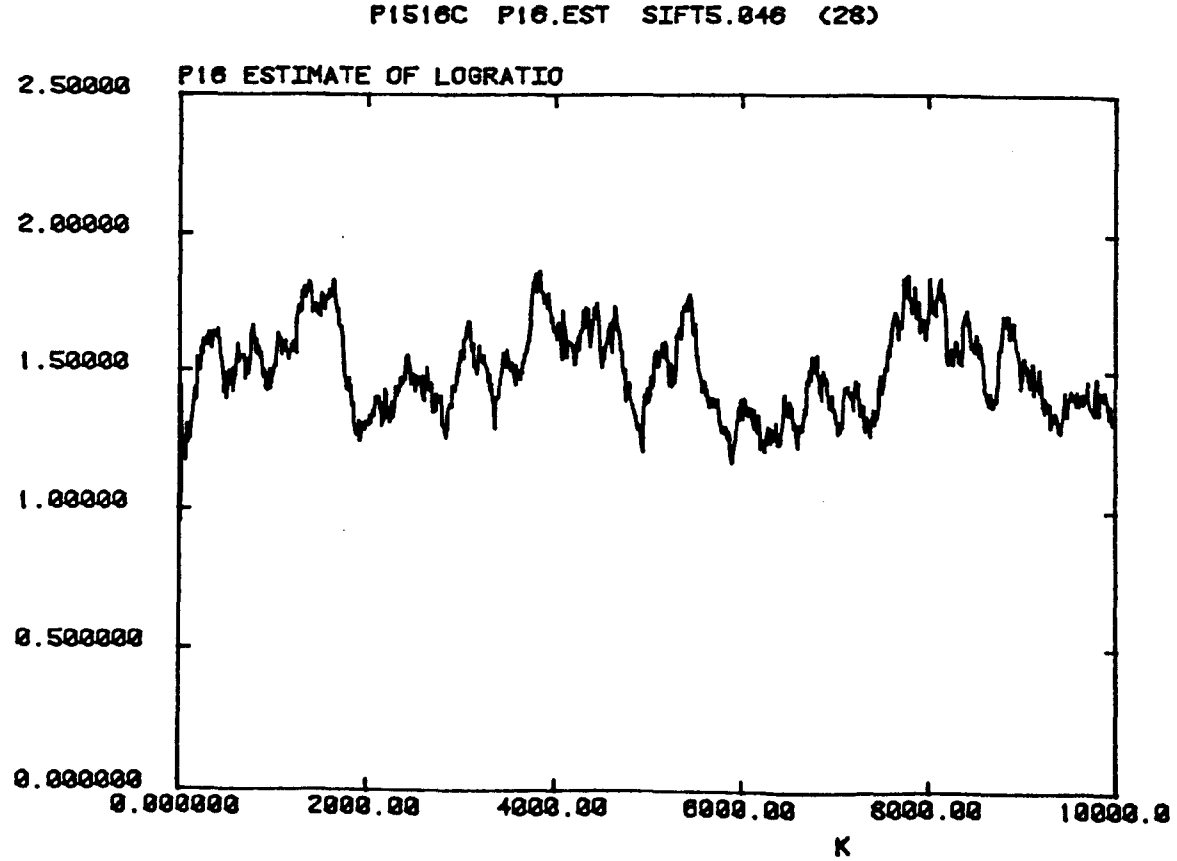
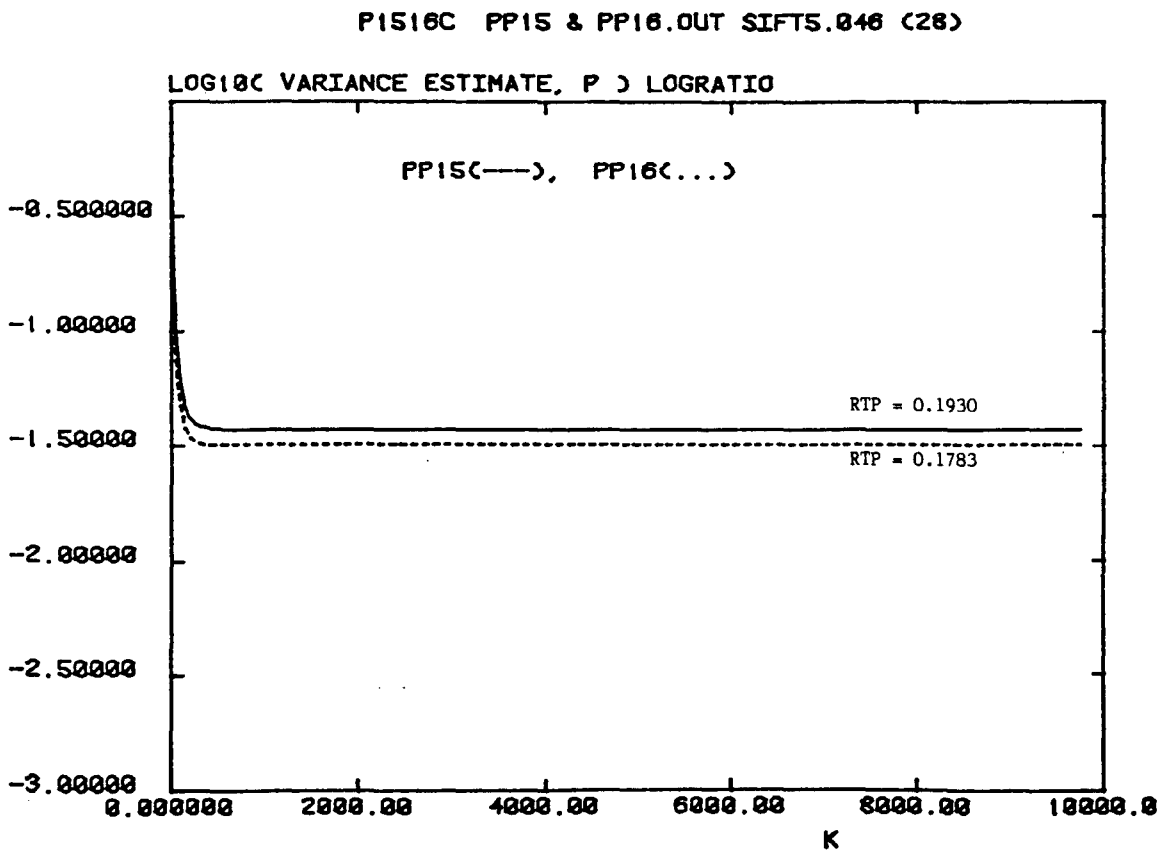


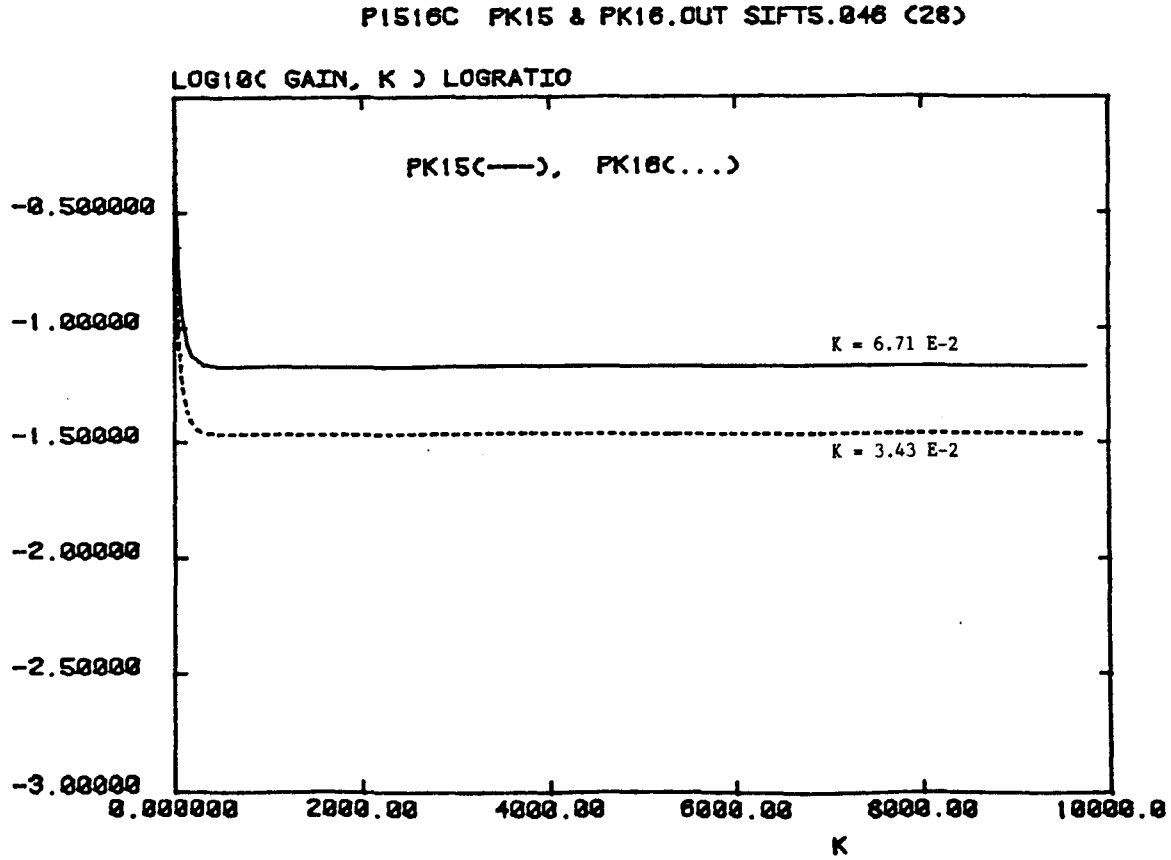
Figure 7.22 (c)



Error Covariance Estimates for SIFT 5·046 (above)

Kalman Gains Used With SIFT 5·046 (below)

Figure 7.22 (d)



Innovation Sequence acfs for SIFT5·046

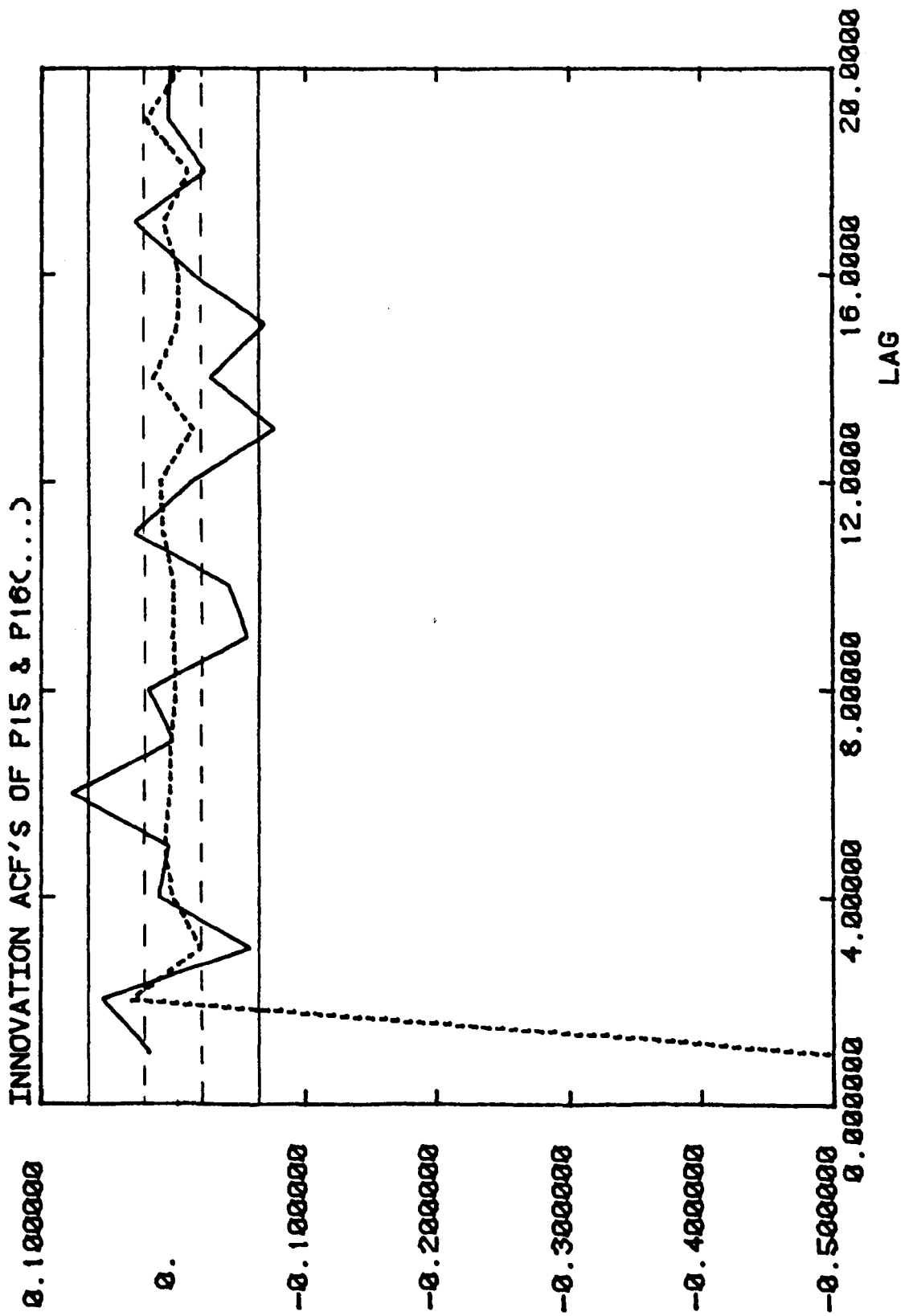
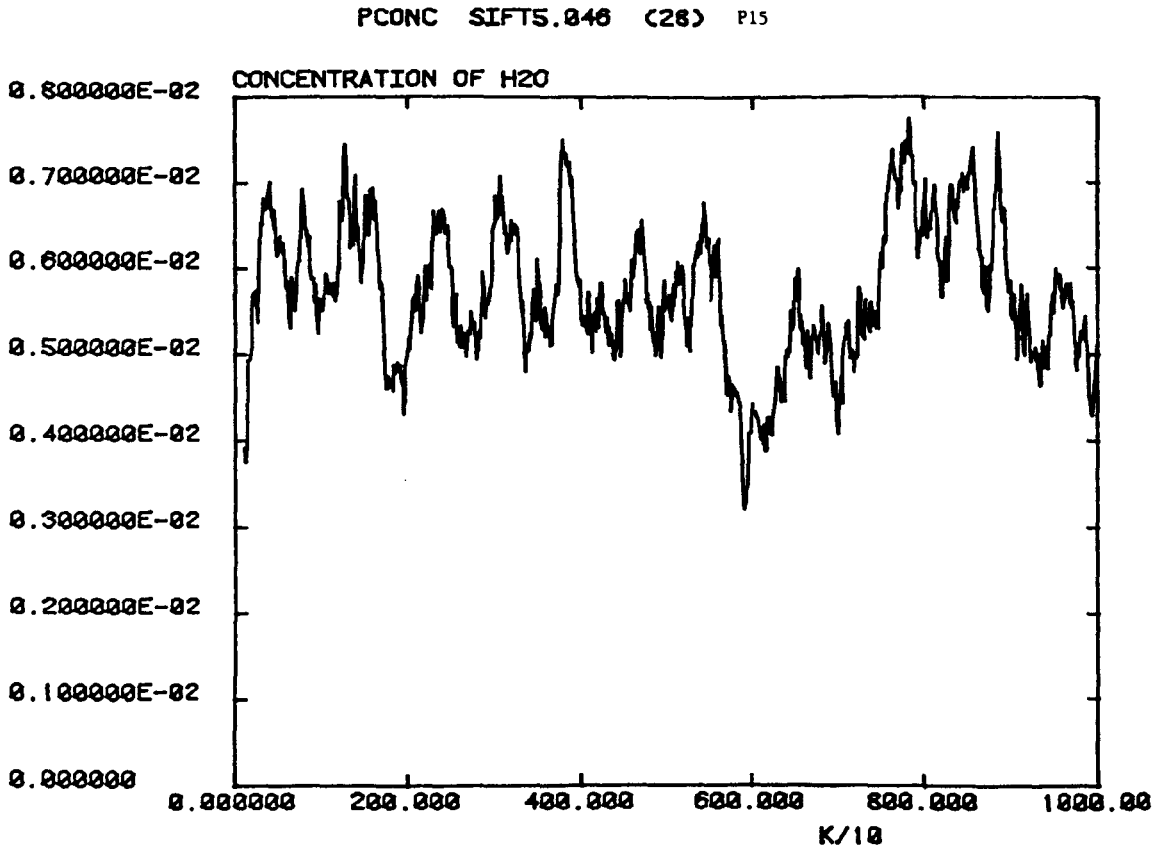


Figure 7.22 (e)

Figure 7.22 (f)



H_2O Concentration Using Log Ratio Estimates of SIFT5·046:
P15 (above)
P16 (below)

Figure 7.22 (g)

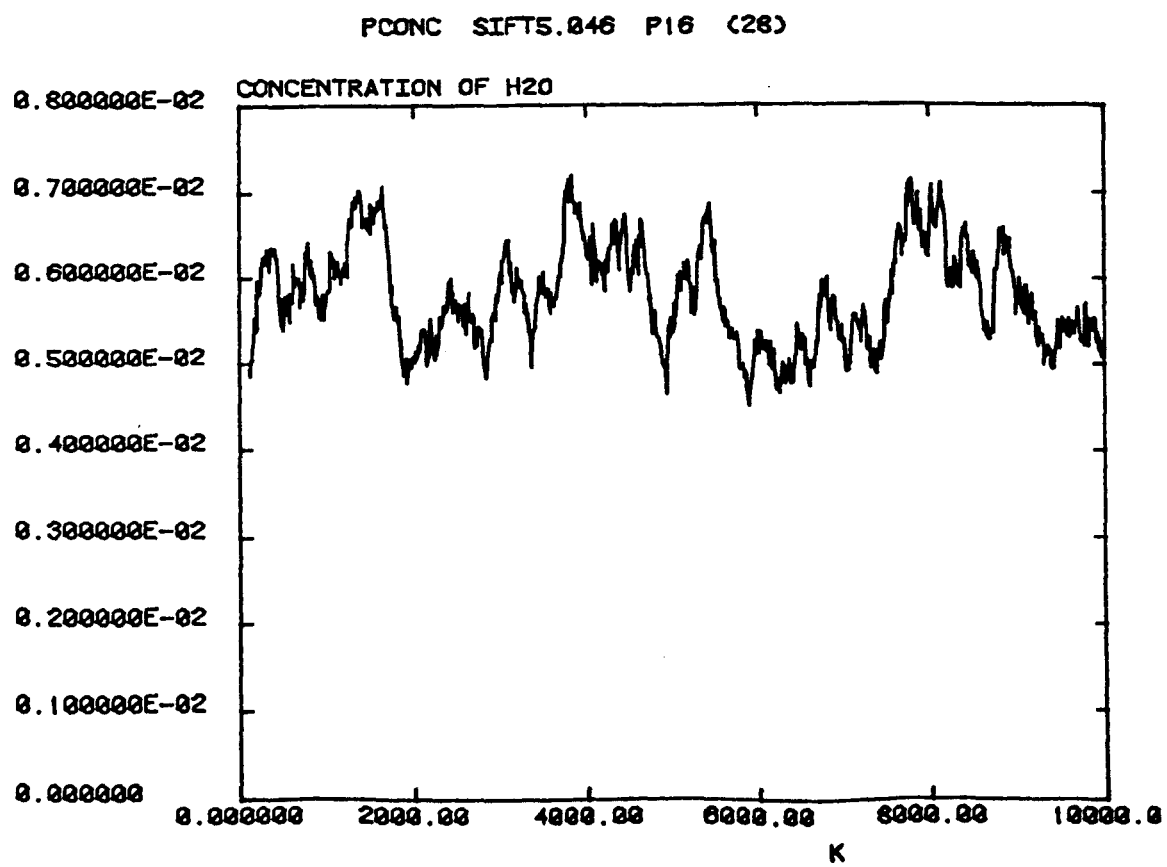
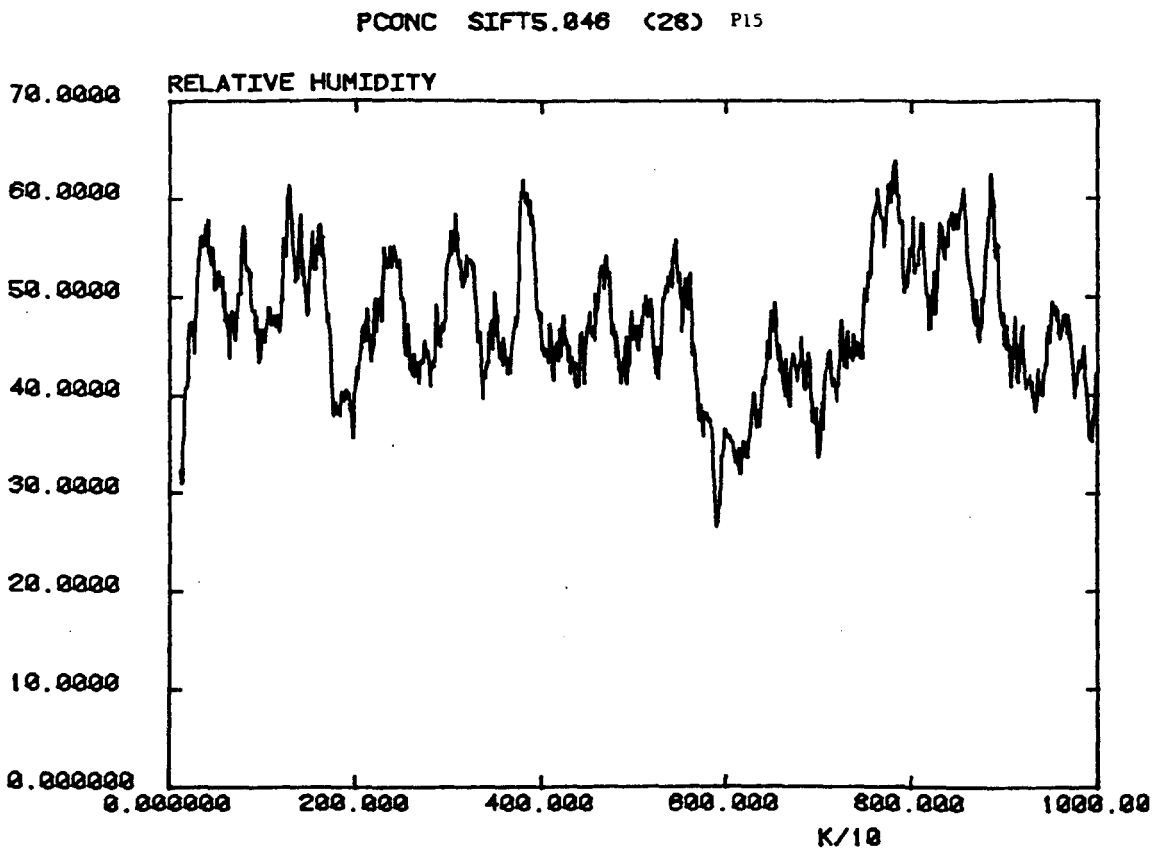


Figure 7.22 (h)



Relative Humidity using Log Ratio Estimates of SIFTS·046:
P15 (above)
P16 (below)

Figure 7.22 (i)

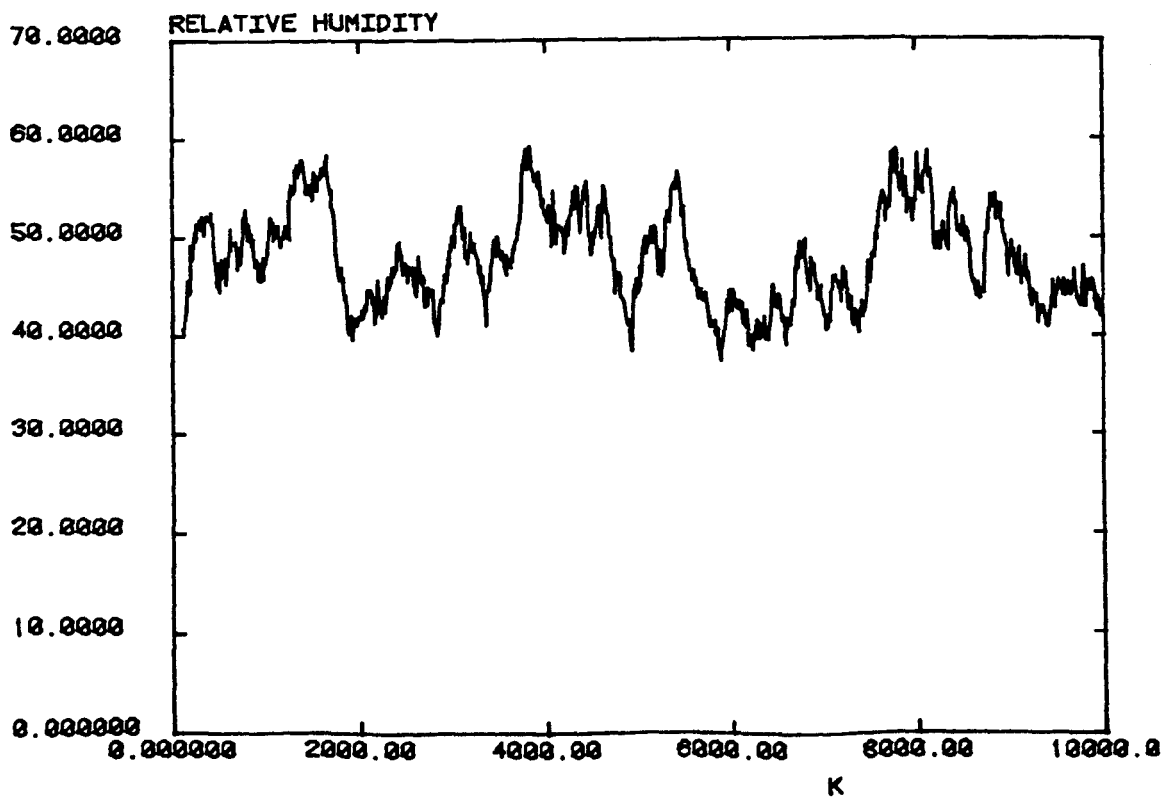
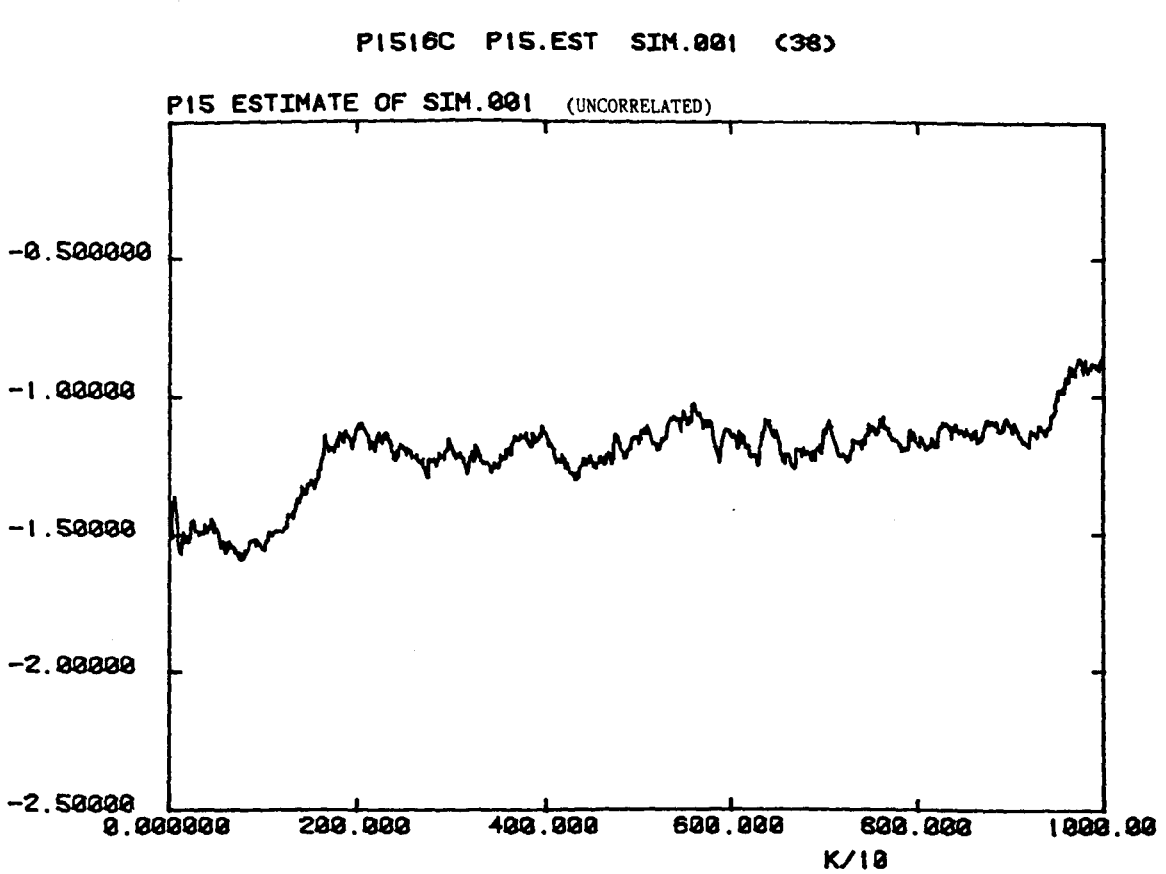


Figure 7.23 (a)



P15 Estimates of Simulated Data (Random Walk Signal):
Uncorrelated Speckle (above)
Correlated Speckle (below)

Figure 7.23 (b)

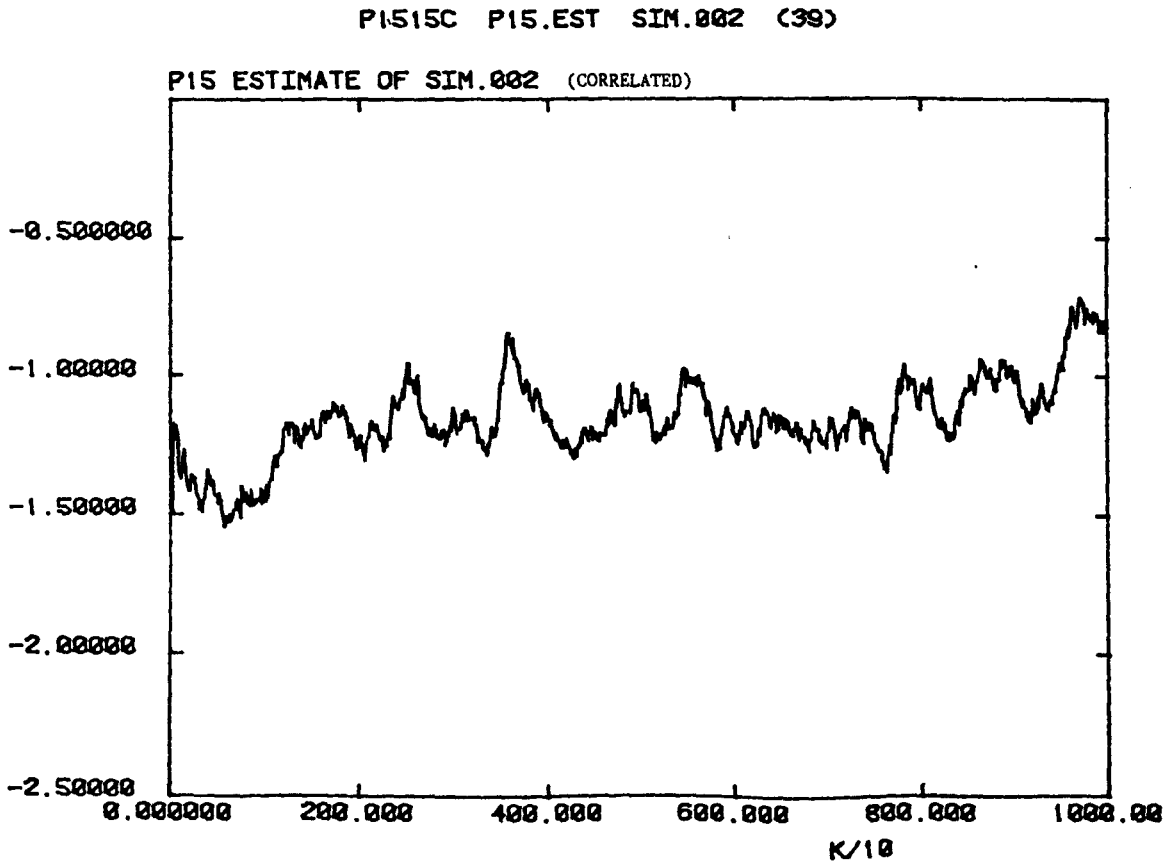
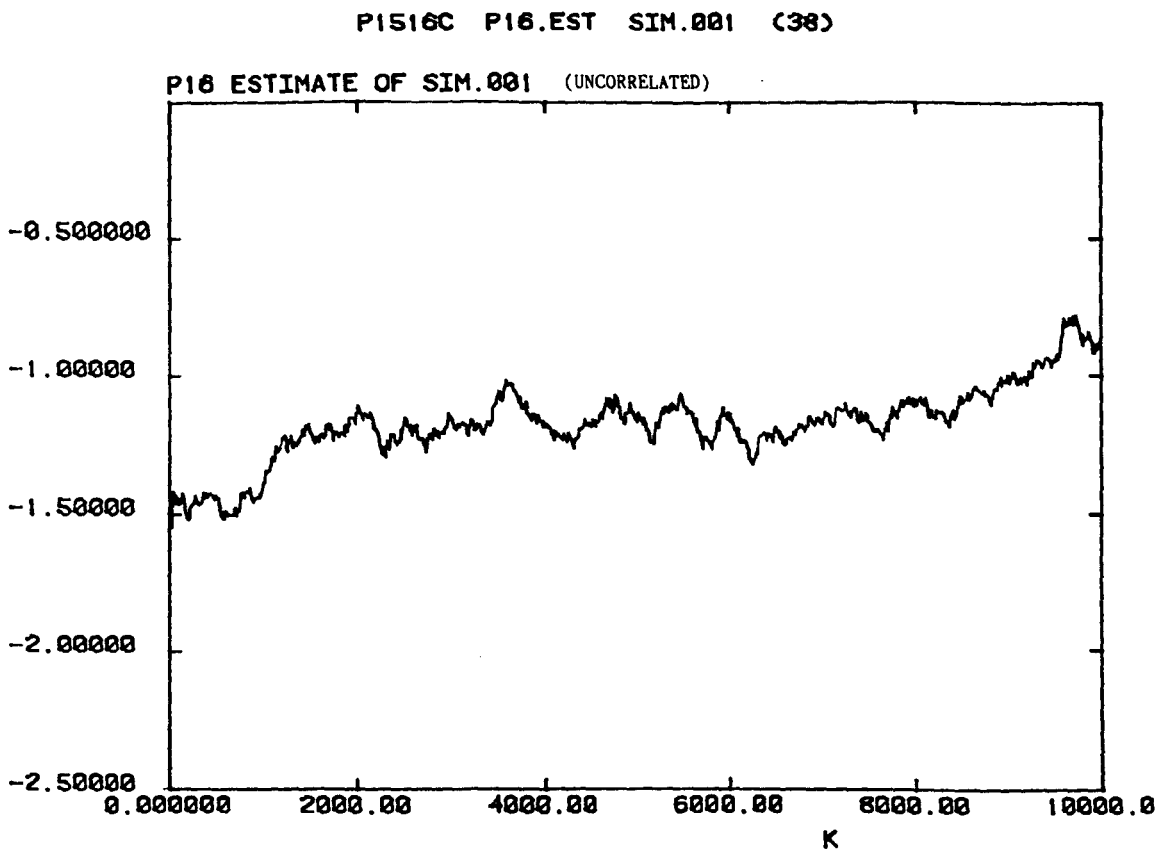


Figure 7.23 (c)



P16 Estimates of Simulated Data (Random Walk Signal):
Uncorrelated Speckle (above)
Correlated Speckle (below)

Figure 7.23 (d)

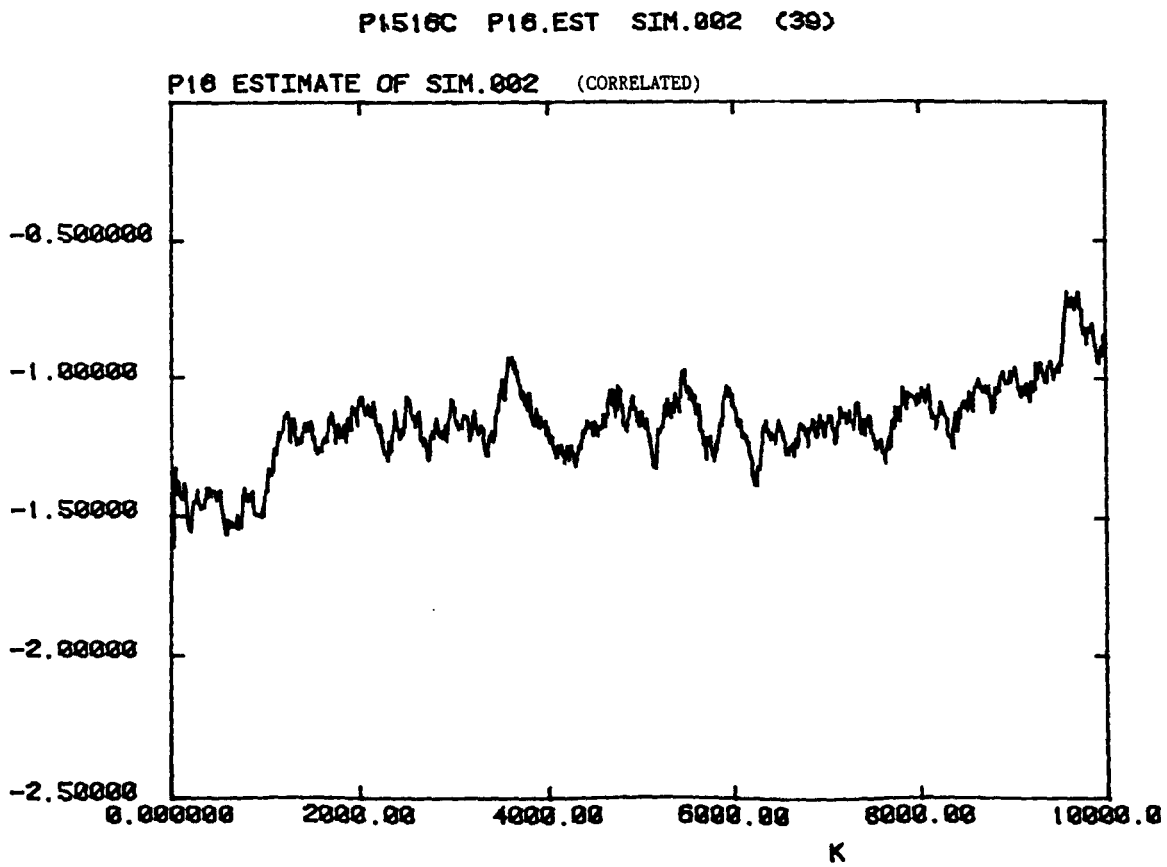
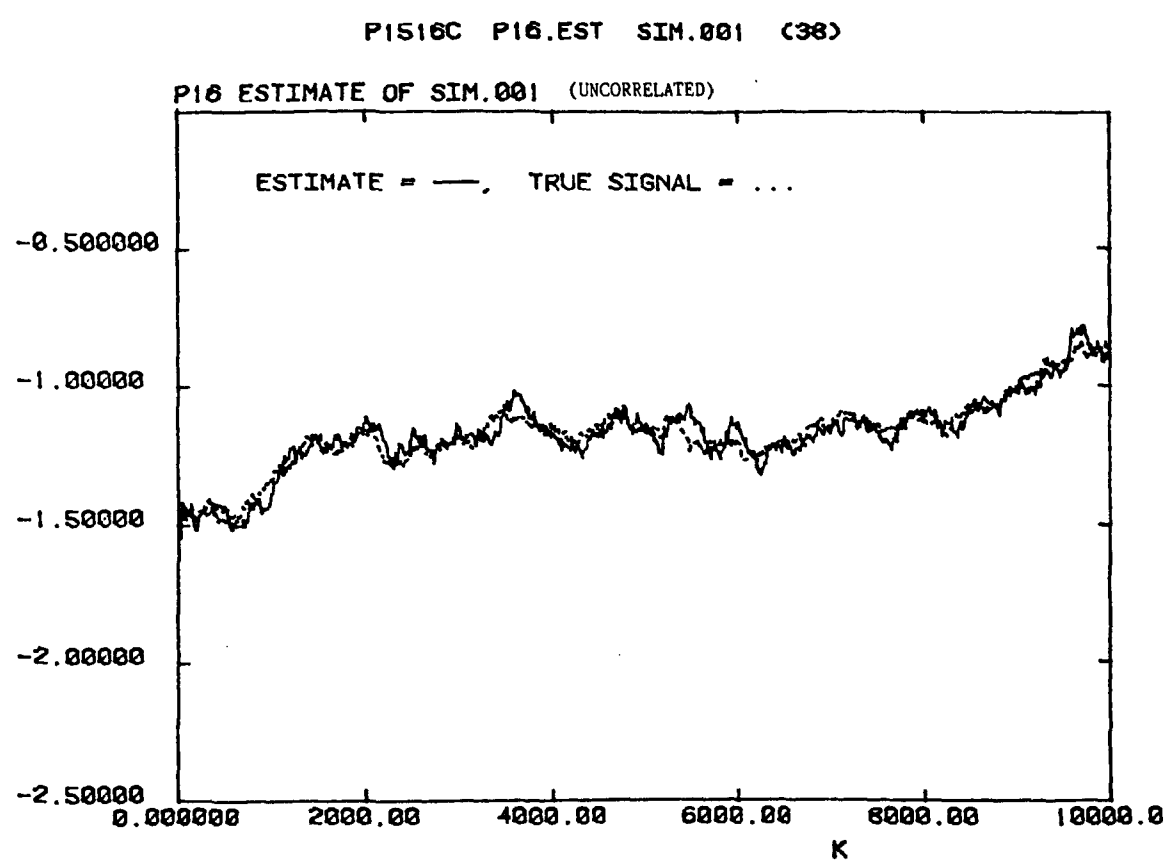


Figure 7.23 (e)



P16 Estimates of Simulated Data (Random Walk Signal):
Uncorrelated Speckle (above)
Correlated Speckle (below)

Figure 7.23 (f)

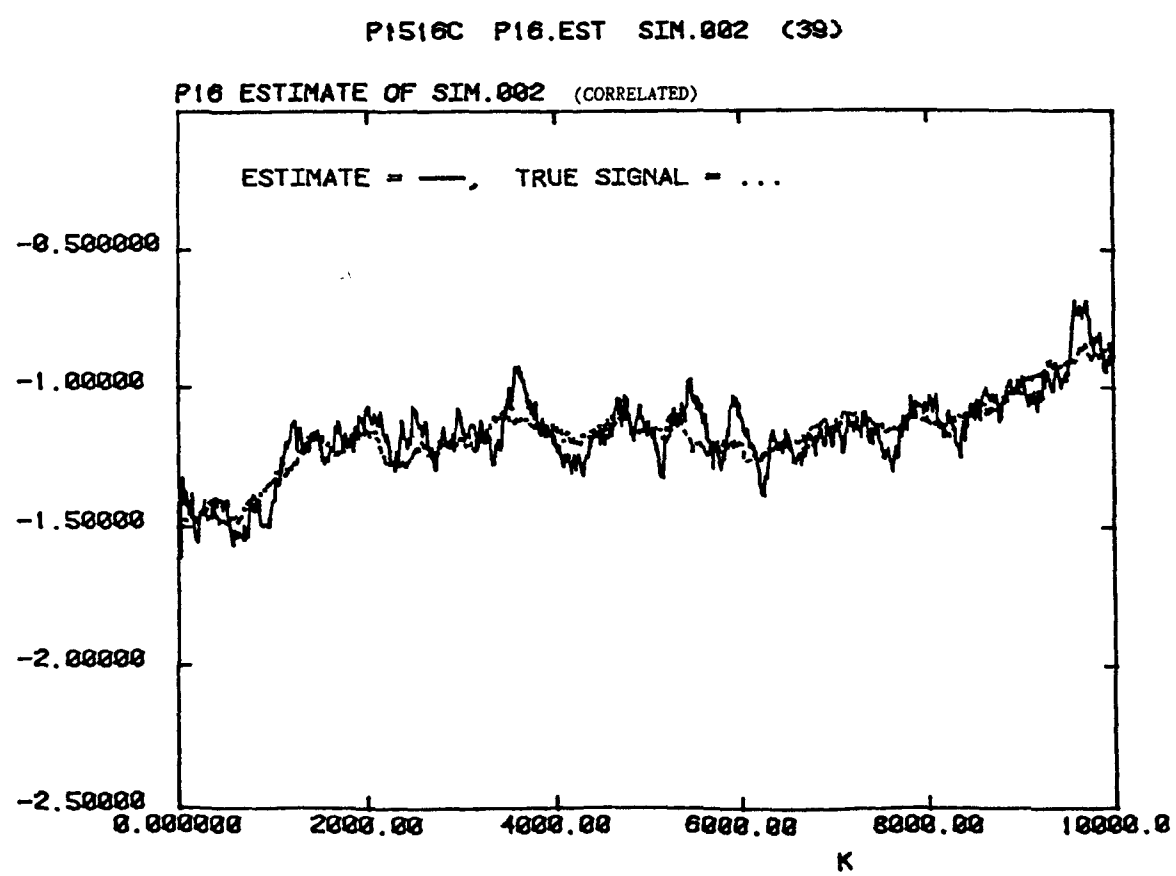
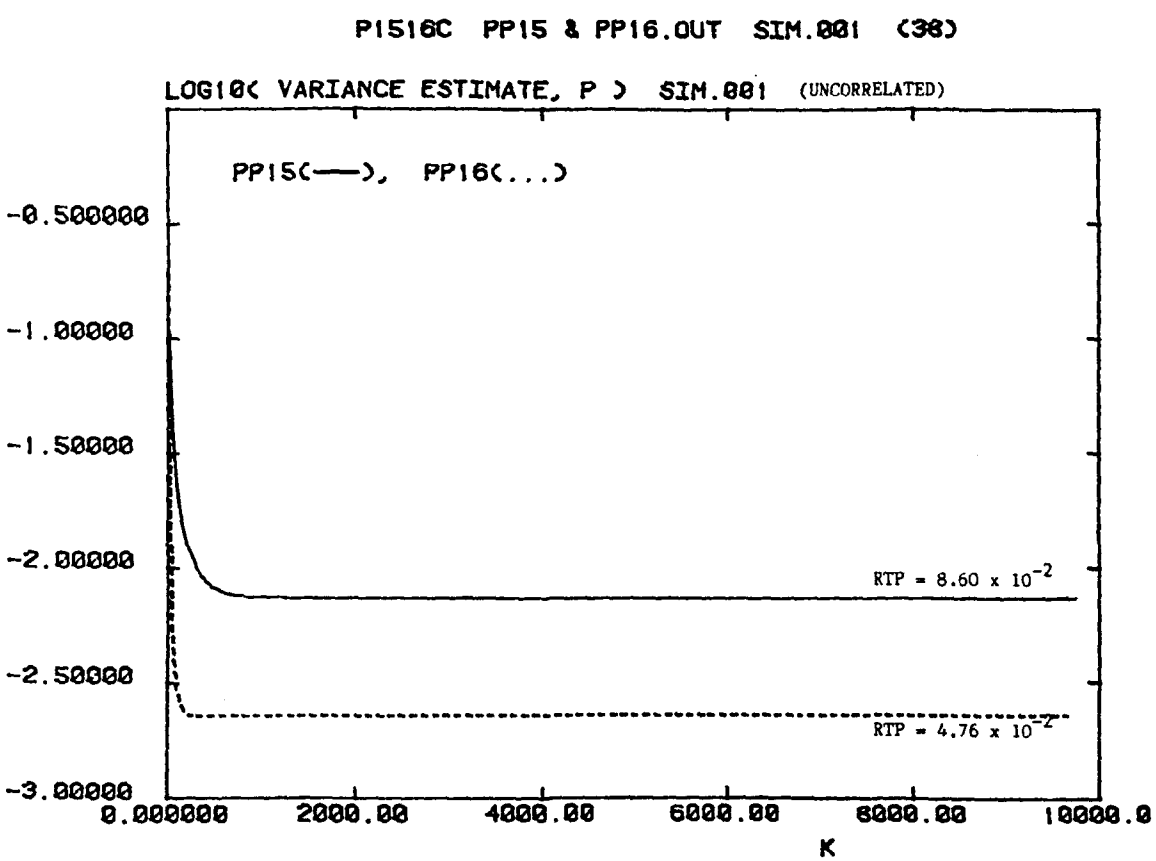


Figure 7.23 (g)



Error Covariance Estimates for Simulated Data:
Uncorrelated Speckle (above)
Correlated Speckle (below)

Figure 7.23 (h)

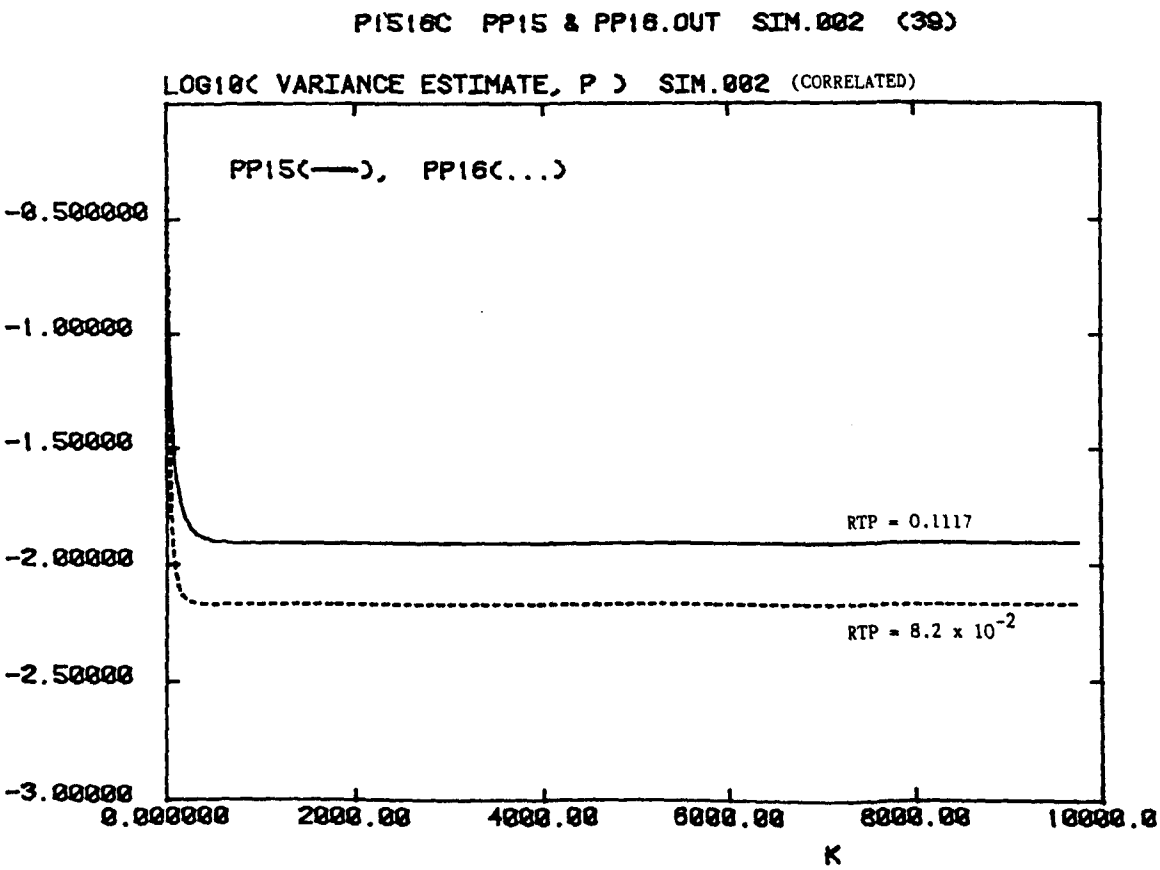
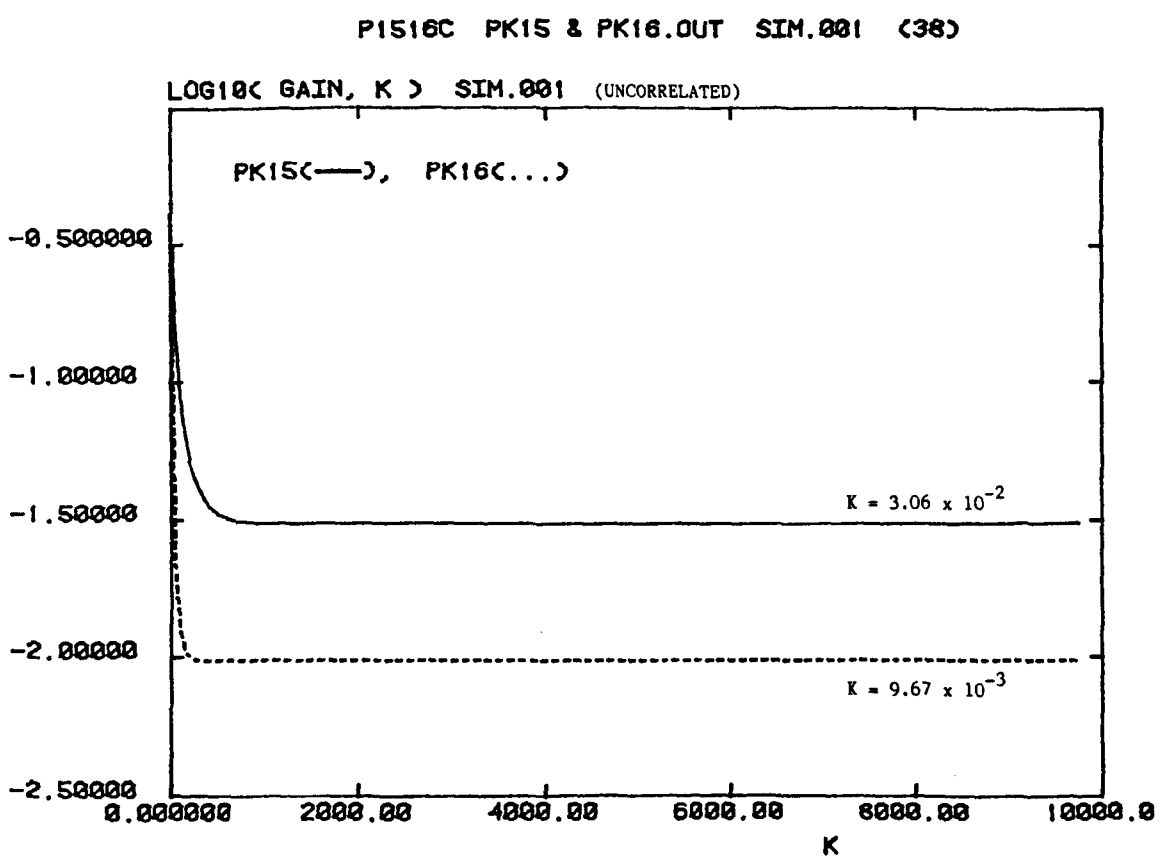


Figure 7.23 (i)



Kalman Gains Used With Simulated Data:
Uncorrelated Speckle (above)
Correlated Speckle (below)

Figure 7.23 (j)

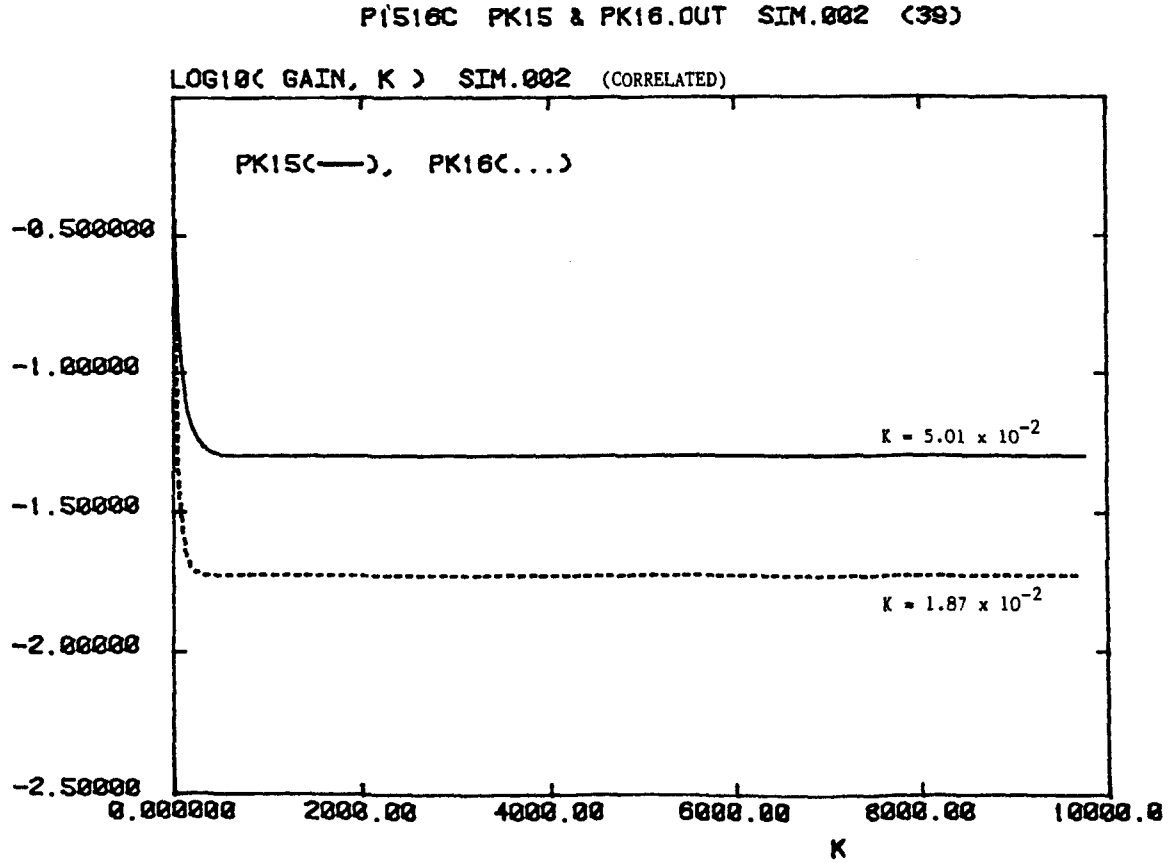
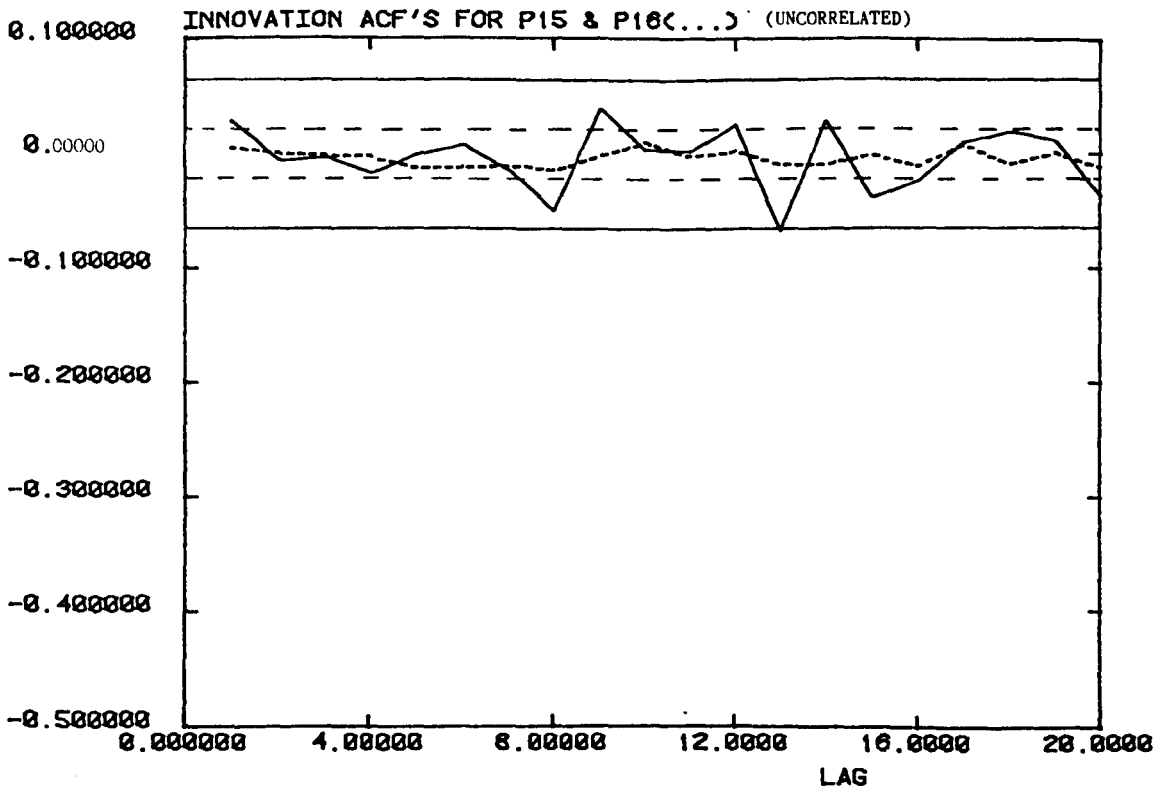


Figure 7.23 (k)

P1516C PACF15 & PACF16 SIM.001 (38)



Innovation Sequence acfs for Simulated Data:
Uncorrelated Speckle (above)
Correlated Speckle (below)

Figure 7.23 (l)

P1516C PACF15 & PACF16 SIM.002 (39)

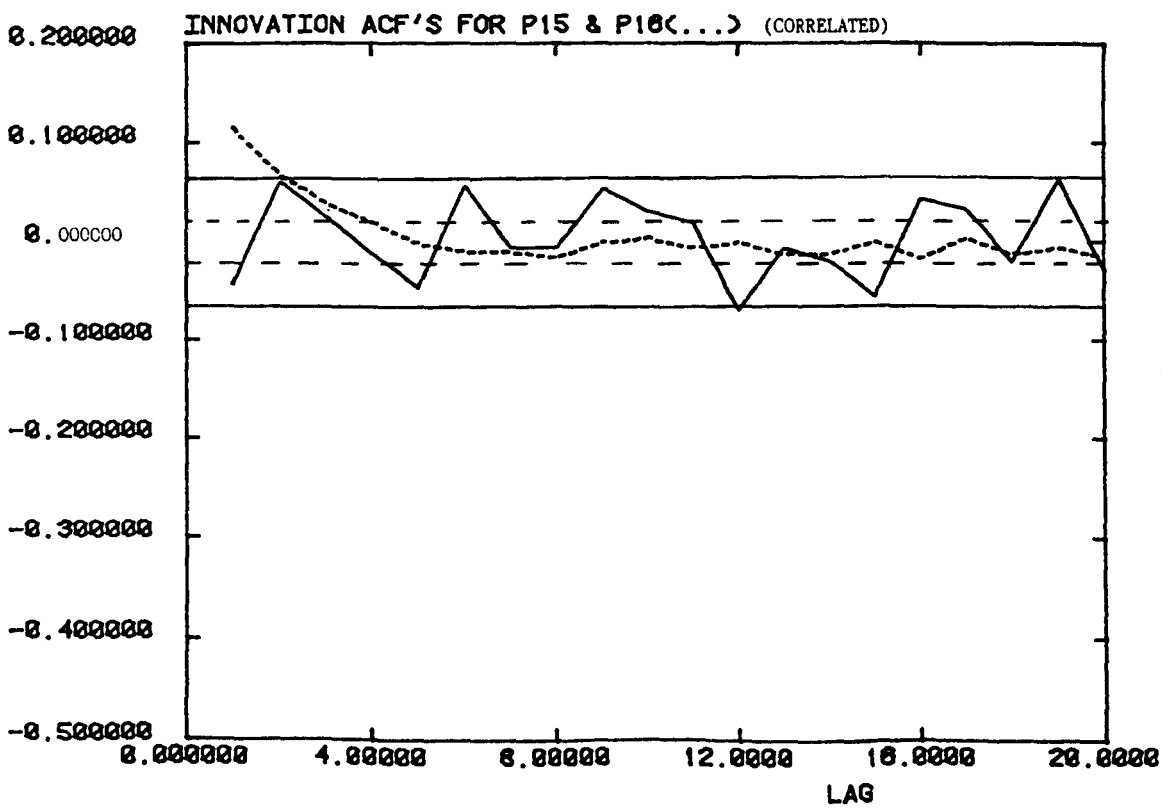
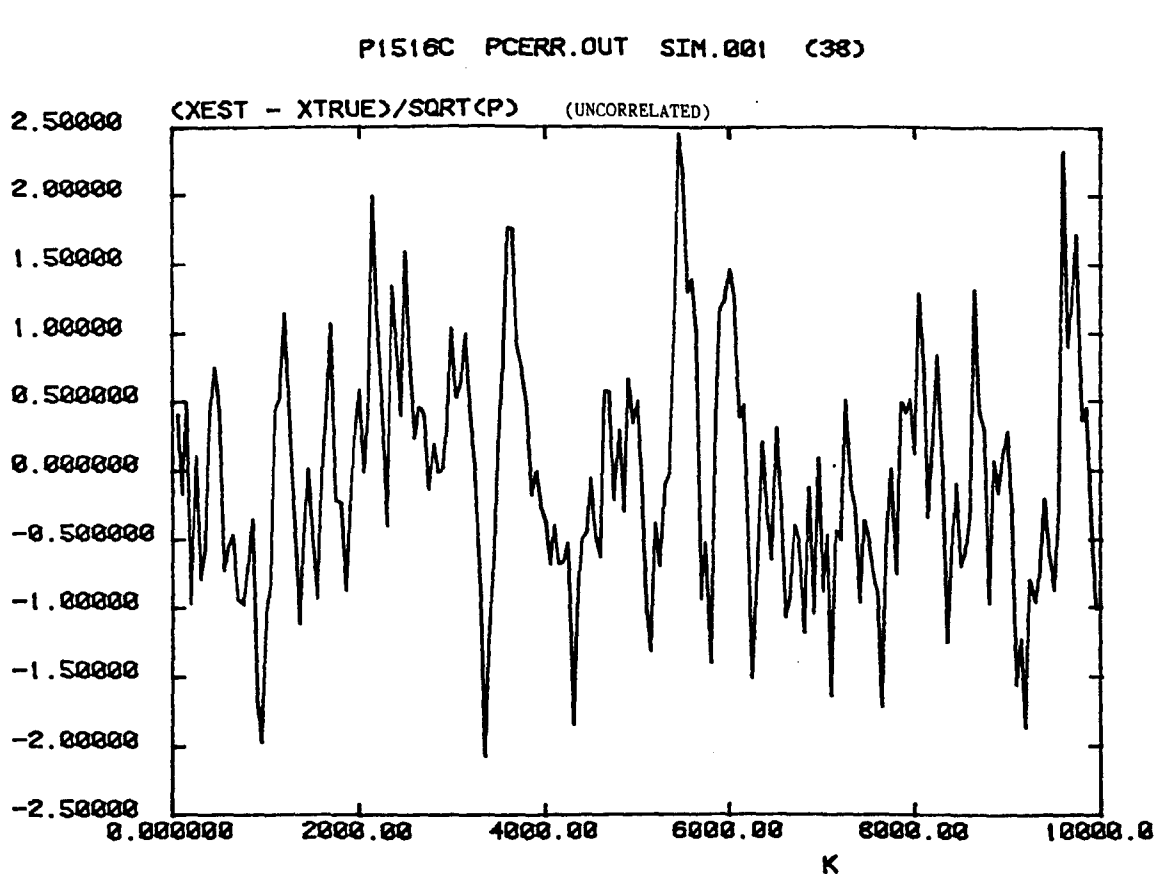


Figure 7.23 (m)



Ratio of True Estimation Error to Estimated Estimation Error:
Uncorrelated Speckle (above)
Correlated Speckle (below)

Figure 7.23 (n)

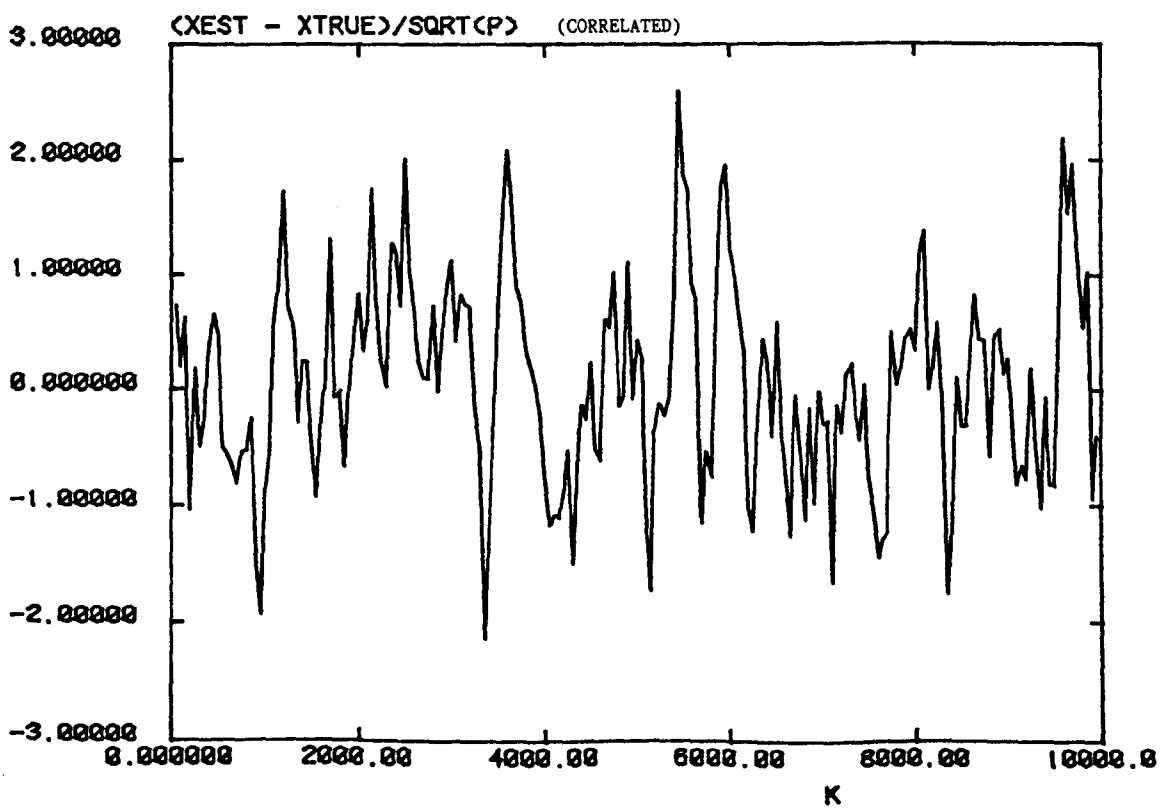
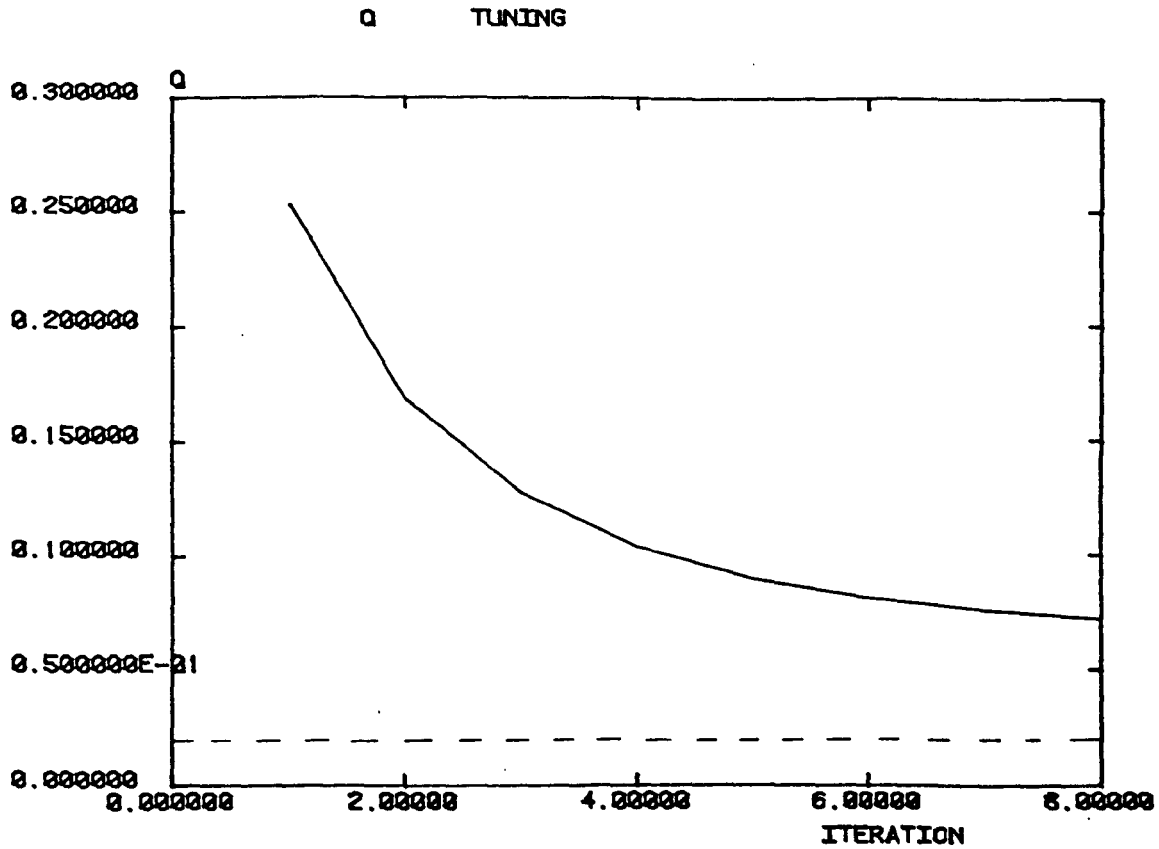
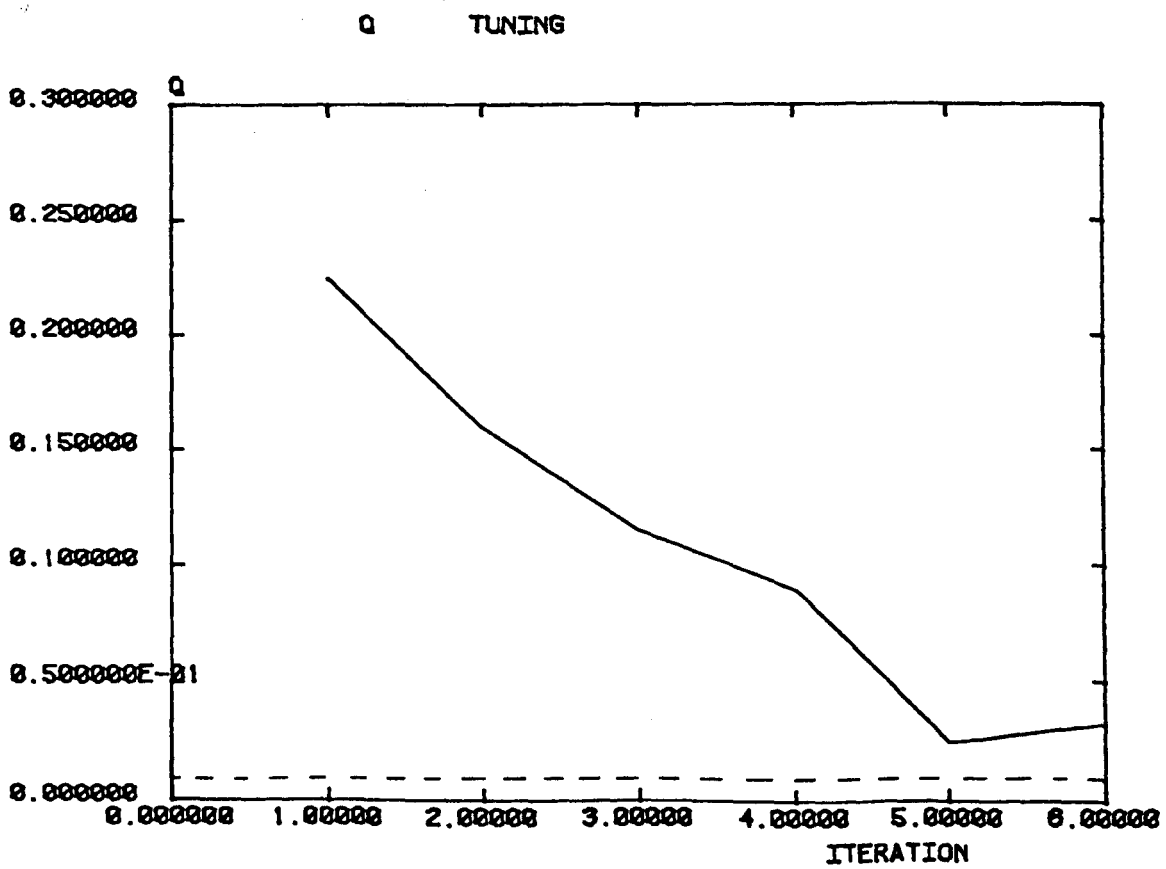


Figure 7.24 (a)



Noise Variance (Q) Estimates as a Function of Filter Iteration Number

Figure 7.24 (b)



Estimation Error as a Function of AR model Coefficient, α.

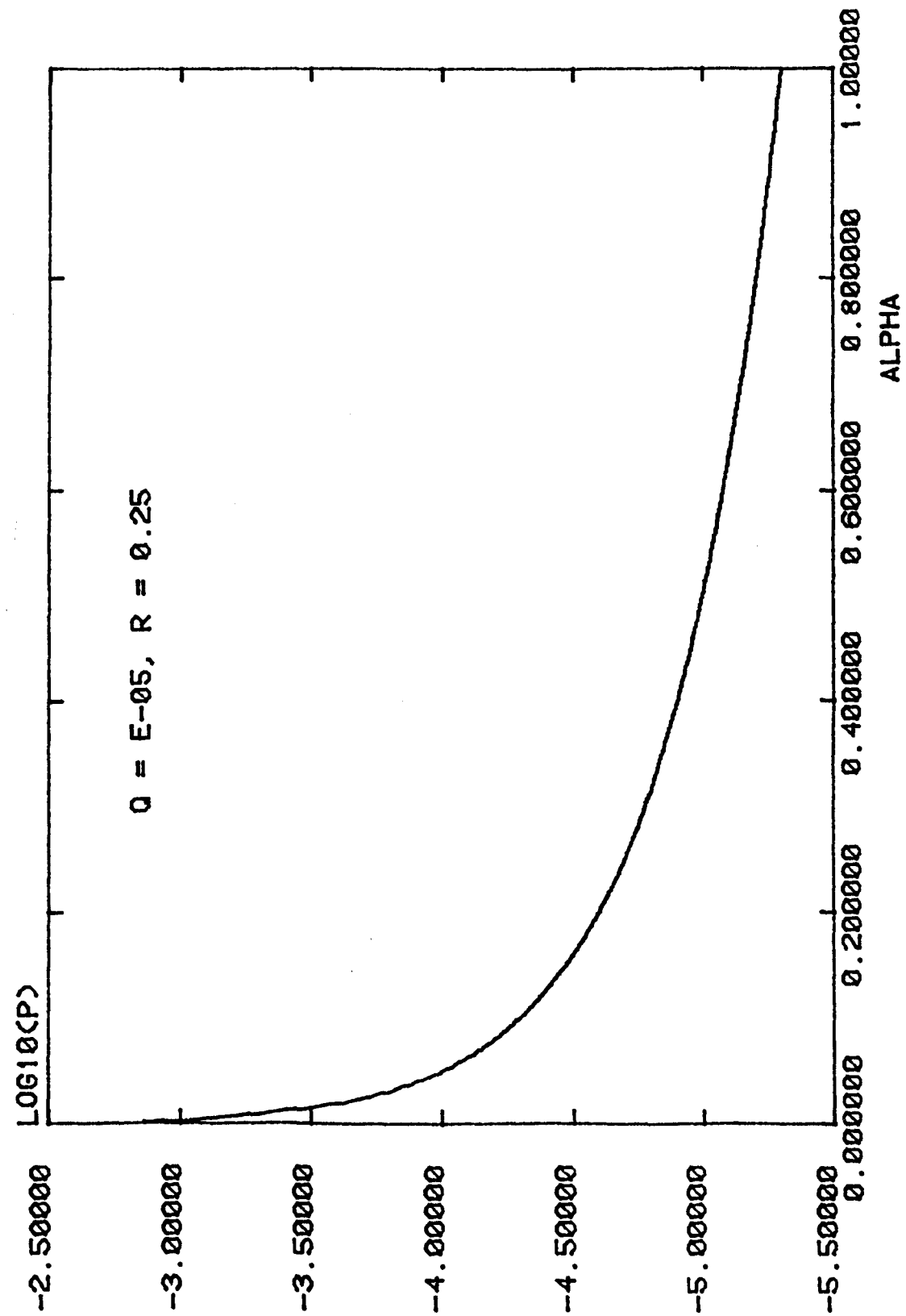
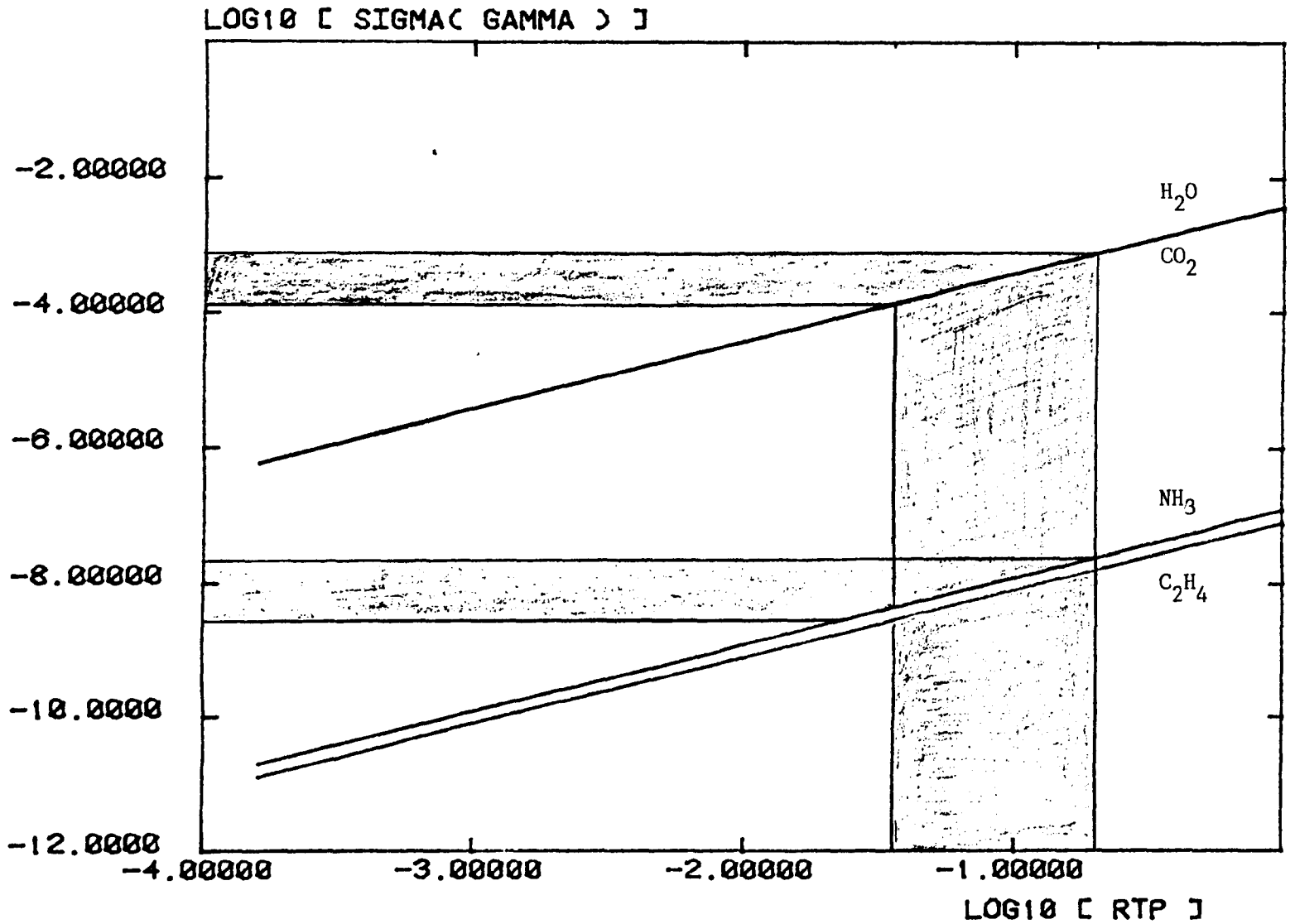


Figure 7.25

PRECISION IN CONCENTRATION

Figure 7.26



CHAPTER 8.0

CONCLUSION

Various estimation techniques have been applied here in an attempt to evaluate their potential for the specific task of processing non-stationary direct detection CO_2 lidar signals corrupted by speckle noise. Two "sub-optimal" algorithms were considered first; the running average and the $\alpha\beta$ tracker, principally as techniques for providing an initial cursory examination of the data; suitable for a qualitative comparison of the information in each lidar channel. The running average is a technique representative of the methods applied by Menyuk and Killinger which first alerted them to the presence of non-stationarity in the signal. The $\alpha\beta$ tracker is a more sophisticated technique, particularly in the slightly underdamped form used here (attributable to Benedict and Bordner), since the algorithm generates an estimate of both the signal and the signal derivative with respect to time. For deterministic signals it is also capable of providing an estimate of the precision of the estimate. Both the running average and $\alpha\beta$ tracker are, of course, essentially deterministic in nature.

Optimal estimation methods require information not only on the statistical properties of the system and measurement noise processes, but also on terms included in the state propagation and observation matrices, ϕ and H. In an attempt to derive this information, specific methods of time series analysis were employed, based on the information content of autocorrelation and partial autocorrelation functions (the "Box-Jenkins" approach). This method proved to be reasonably successful, at least for the purposes of selecting relatively simple, first order autoregressive models as approximations to both the signal process and

sequential correlation in the additive speckle measurement noise (after the log transformation).

Kalman filter applications, in general, may be broadly partitioned into one of two problem-type categories. The first, and most widely relevant group would include those applications in which the system model tends to be either deterministic, or else has well known statistical properties, and therefore facilitates the estimation of large state vectors. The other application area is characterised by systems which are probabilistic rather than deterministic, having poorly defined stochastic properties and, as a partial consequence of this, state vectors which tend to be limited in dimension. Lidar signal estimation, as attempted here, currently falls into the latter category.

Adopting first order autoregressive models in the Kalman filter implementations of Chapter 7 (P15 and P16), simplified an initial application of two complementary techniques which may be used as part of an iterative procedure for improving parameter definition within the models:

(1) The first technique addresses the immediate requirement to evaluate the system and measurement noise covariance matrices, Q and R. Reducing the system and measurement models to scalar form makes this easier since only two scalar quantities, q and r, need to be determined. The method published by Mehra is a two stage process. Initially, crude estimates of q and r are used in a standard implementation of the Kalman filter (i.e. P15) and the autocorrelation function of the innovation sequence is calculated. A statistical test is then used to determine whether or not the innovative sequence is white.

If it is not, the second step is to use Mehra's formulations for deriving new estimates of q and r. Those are then used in a re-run of the filter and the procedure repeated until the whiteness test is successful.

(2) The basic, linear Kalman filter does not provide for the eventuality of correlated measurement noise; a significant limitation since measurements of real processes, in general, will be corrupted by noise which must be band limited in some sense. Bryson and Hendrikson's algorithm is therefore a valuable extension to the linear Kalman filter but since at least one extra model parameter (Ψ or β) must be estimated, the implementation requirements become more complex and filter performance will be susceptible to at least one extra source of error.

Other algorithms for estimating q and r, or K, have been published since Mehra's paper which may be capable of providing more accurate estimates of these parameters, and would therefore benefit from a comparative analysis based on both real and simulated lidar data. True adaptive filtering, in which q and r, and consequently P and K, are estimated "on line" could then be attempted using the best of these methods.

Where, in future applications, more complex models become necessary, the Box-Jenkins approach may be pursued further to identify the coefficients of a general ARMA process. Alternatively, if parameter instability is suspected, the "Instrumental Variable" methods advocated by Peter Young [1] may prove to be more appropriate. Certain topics which are generally relevant when applying the methods of optimal estimation have not been discussed here because of the lengthy digres-

sions that would have been involved. These include the subjects of smoothing, sensitivity analysis and word length errors.

Smoothing, as defined at the beginning of Chapter 6, is a non-real-time processing scheme that uses all of the measurements before and after each sample, Z_k, to estimate the system state, instead of just those measurements up to and including Z_k, as is the case with filtering. Had this been done, the steady state error covariances for the random walk system model could have been reduced by a factor of 2. However, once the filter is defined, it is relatively simple to implement the smoothing version of the algorithm.

Sensitivity analysis examines state estimate sensitivity to changes in the model parameters, and usually refers to the situation in which the system and measurement processes are known precisely, but a simpler, approximate implementation is required to ease computational loading. In a general sense the term can also be applied to an analysis based, for example, only on the error covariance estimates. Gelb [2], in fact, gives an example for a first order gauss-markov system model (AR(1) with $\phi < 1$). In the limit in which this becomes a random walk sensitivity analysis, in the broad sense, would then reduce to the simple task of examining filter performance as a function of r and q.

Word length errors are a consequence of using digitised computers in which numerical precision is a function of the number of bits used to represent a number. During a multiplication, for example, two numbers represented by n bits each may require 2n bits for their product. In fixed word-length machines only n bits are available, however, so the precision represented by the last n bits must be discarded. There are two methods of doing this; symmetric rounding up or down, or truncation to the nearest lower integer. The latter technique results

in the largest accumulation of errors, particularly during the multiply-accumulate operations characteristic of recursive filtering. A check on the Fortran implementation of the filtering algorithms was conducted by converting all susceptible variables to double precision and repeating several filter runs on real and simulated data. No differences were observed due, presumably, to symmetric rounding techniques used in the Fortran floating point library routines.

In applying data processing techniques to the problem of speckle induced measurement noise the alternative approach, of improving the sampling capabilities of the instrumentation used to collect the data, has been neglected. As indicated in the discussion on speckle in Chapter 4, the instrumentational requirements translate into a need to reduce the ratio of the speckle induced standard deviation to the signal mean, given by $1/\sqrt{m}$. In principle, uncorrelated speckle patterns can be obtained by means of time, space, frequency or polarisation diversity. For example, if the diameter of the receiver mirror is increased, equation (4.15) reveals that the size of the speckles in the image plane will be decreased. More speckles will therefore be averaged over a given detector dimension since m is increased; a technique which is referred to as an aperture averaging. Alternatively, for fixed receiver diameters, the location of the beam on the target may be deliberately altered between shots to decorrelate successive power samples.

Advanced data processing techniques will, however, find application where instrumental techniques are either impractical or too costly to implement. CO_2 lidar systems, in general, have yet to become the ubiquitous tool for atmospheric gas concentration measurements they originally promised to be. This is attributable, principally, to their long and expensive development phases which are often dependent on

advances in laser technology. Digital filtering methods may therefore provide at least an intermediate solution, and are now capable of real time implementation via relatively inexpensive, VLSI based, programmable digital signal processing systems.

References

[1] Young, P.; "Recursive Estimation and Time Series Analysis: An Introduction", Springer Verlag, 1984

[2] Gelb, A. (Ed), "Applied Optimal Estimation", The M.I.T. Press, 1974.