AD-A083 645   SYSTEMS CONTROL INC PALO ALTO CA   F/G 9/4
IDENTIFICATION OF DYNAMICAL SYSTEMS IN THE PRESENCE OF NON-GAUS--ETC(U)
DEC 79 H SALZWEDEL, N K GUPTA, W E HALL   N00014-78-C-0519
UNCLASSIFIED SCI-TR-5274-04   NL

1 OF 1
AD A
083645

END
DATE
FILMED
6-80
DTIC

(2)

SYSTEMS CONTROL, INC. (Vt)
1801 Page Mill Road
Palo Alto, California 94304

Telex: 348433

Telephone:
(415) 494-1165

(11) DEC [redacted] 79

14 SCI-TR-5274-04

12 63

(16) IDENTIFICATION OF DYNAMICAL SYSTEMS IN THE PRESENCE OF NON-GAUSSIAN AND NON-WHITE NOISE.

(9)

Prepared by

(10)
H. Salzwedel
N.K. Gupta
W.E. Hall

Prepared for

OFFICE OF NAVAL RESEARCH
800 North Quincy Road
Arlington, Virginia 22217

80 3 27 039

FOREWORD

The work reported here was sponsored by the Office of Naval Research under the direction of Mr. Robert Van Heusen and Mr. David Siegel. The research was conducted under Contract N00014-78-C-0519 between November 1978 and November 1979.

At Systems Control, Inc. (Vt), N.K. Gupta was the project manager and W.E. Hall, the program manager. H. Salzwedel was the project engineer.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

I. INTRODUCTION & SUMMARY

1.1 INTRODUCTION

Systematic procedures for the identification of dynamic systems have been developed over the last two decades. These methods have been successfully applied to a variety of vehicles and other systems. Most of the methods are based on several significant assumptions.

(1) Models used in parameter estimation step are correct.

(2) The state and measurement noise follows a Gaussian distribution (this assumption is made both in model structure determination and parameter identification).

(3) The noise sources are white or have a known rational spectrum.

(4) All unknown parameters about which there is information in the data are identified.

(5) Sufficient data is available such that asymptotic estimator properties are valid.

Real data often do not follow these assumptions leading to an inefficient estimator. The following symptoms which indicate a lack of estimator efficiency have been observed.

(1) Residuals are non-white and non-Gaussian.

(2) The actual estimation errors are much higher than those predicted by statistical analysis (Cramer-Rao bounds).

(3) The estimation errors are unacceptable when too many parameters are estimated.

(4) The parameter estimates often have smaller error when zero state noise is used compared to the estimates when true state noise is used.

(5) It is extremely difficult to get good results from short data records.

System identification methods are at a stage where the issues described above need to be attacked. This report formulates procedures to treat non-Gaussian, non-white noise statistics in order to develop systematic algorithms and an interpretative framework for treating actual data.

1.2 RESULTS

The investigation into nonconventional noise sources has been divided into two parts. The first part develops parameter estimation methods with non-Gaussian noise in state and measurement equations. The following is a summary of significant results of this work.

(1) Heavy tailed distributions can markedly degrade estimation accuracy. Thin tailed or amplitude limited noise has minor influence on estimation error.

(2) A simple rejection of outliers approach is both statistically inefficient and computationally undesirable.

(3) Advanced methods are developed that lead to improvements in both state and parameter estimation accuracy. (Thus, a robust Kalman filter is a byproduct).

(4) A simple example demonstrates that parameter estimation errors can be reduced by a factor of three by using robust estimation. A larger improvement is expected in parameters which are only marginally identifiable.

The second part of the work studies non-white noise. The following results have been obtained.

(1) The non-white noise does not cause a bias in estimates.

(2) Low frequency noise usually increases estimation error while the high frequency noise decreases it. Unfortunately, most systems have low frequency noise.

(3) The non-white noise could be corrected by building an appropriate filter or a whiteness insensitive estimator can be designed.

(4) Cramer-Rao bounds based on white noise assumptions are significantly different than if a colored noise assumption is used. Since most current analyses are

based on the the white noise assumption, Cramer-Rao bounds have been a poor measure of estimation errors. The noise spectrum should be estimated and the correct spectrum should be used in deriving estimates of errors.

(5) When the noise spectrum is nonrational (not a ratio of polynomials), optimal parameter estimators are difficult. It is often reasonable to approximate the spectrum by a ratio of polynomials.

1.3 SUMMARY

Section II of this report describes parameter estimation problems associated with non-Gaussian noise. This is followed by the discussion of non-white noise in Section III. Finally, conclusions and areas of future investigation are given in Section IV.

II. PARAMETER ESTIMATION IN THE PRESENCE OF NON-GAUSSIAN NOISE

2.1 INTRODUCTION

Measurement errors are the sum of inaccuracies from a number of sources. These errors can be divided into two broad classes: (1) systematic errors, and (2) random errors. Each error follows some probability distribution but is otherwise unpredictable. The systematic errors are identified during the measurement system calibration tests. During the parameter estimation stage, these errors are set to test values or are jointly estimated with states/parameters of interest. Since random errors change with time in an unpredictable manner, their effect is minimized by the use of a filter.

For measurements z, dependent on parameters θ, the most likely values $\hat{\theta}$ of the parameters θ can be determined by least squares, minimum variance, or maximum likelihood methods. All these methods assume that noise probability distributions are known and all errors follow the assumed probability distributions. The Gaussian assumption, for example, leads to the least-squares solution. Approximations are required to provide practical solutions in estimation problems. Failures of components in instrument systems, local inaccuracies, sudden environmental changes, and the occurrence of gross errors are normally not considered in the assumed probability distribution of the measurement system noise parameters. Because of largely increased complexity in modern sensor systems, however, these errors have become increasingly more important and define the need for estimation procedures that are not very sensitive to departures from the assumptions on which they depend, robust estimation.

2.2 TREATMENT OF NON-GAUSSIAN NOISE

To treat the subject of non-Gaussian noise sources, two cases are considered: (1) the noise distribution is known, and (2) the noise distribution is not known.

When the noise distribution is known, maximum likelihood, minimum variance or Bayes' approaches may be used for parameter estimation. Such approaches give nonlinear estimators even for linear systems. In general, optimal estimators for nonlinear systems are complex. To illustrate the likelihood method, Appendix A derives a maximum likelihood parameter estimation algorithm for Poisson noise distribution.

When the noise distribution is not known, two approaches may be used for estimation and identification.

(1) In the first approach, noise probability density functions are estimated. These probability density functions are used to derive optimal estimators.

(2) In the second approach, robust methods are used. These methods are minimally sensitive to distributions of noise. Some efficiency (e.g. accuracy) is lost if the distribution is, in fact, Gaussian.

The rest of this chapter will deal with robust estimators.

2.3 REVIEW OF PREVIOUS WORK

The significance of the Gaussian assumption has been known for a long time. Tukey (1960) and Huber (1972) compared variances of parameter estimates computed by minimizing mean deviations

$$d_n = \arg \min_{\bar{x}} \left(\frac{1}{n} \sum_i |x_i - \bar{x}|\right) \tag{2.1}$$

to those computed by minimizing mean square deviations

$$s_n = \arg \min_{\bar{x}} \left(\frac{1}{n} \sum_i (x_i - \bar{x})^2\right)^{\frac{1}{2}} \tag{2.2}$$

for an error which is normally distributed but is contaminated by another normally distributed random variable whose mean square value is three times higher. They showed that a contamination of 0.2% suffices to make the asymptotic efficiency of the mean deviation larger than the asymptotic efficiency of the mean square deviation (Huber, 1977, p.2). Thus, some estimation methods are very sensitive to deviations from the assumed distribution. Note that the mean deviations method is not the best for mixtures of distributions or distribution uncertainties. Maximum likelihood methods, which fit smooth distribution functions through the actual error statistics (Hall/Gupta, 1974) give generally much better results.

One of the main reasons for the sensitivity of the estimation methods to deviations from assumed error statistics results from their extreme behavior in the distribution tails (Huber, 1972; Hampel, 1971). Pierce (1852), Freedman (1966), and others showed that engineering and scientific data have typically outliers of several percent that fatten the tails of the distribution of random variables. Test data of submarines and aircraft have typically 1% to 5% outliers. Legendre (1805) suggested robustifying the least-squares estimator by rejecting all data which have obvious errors much larger than the remaining data. Airy (1856) pointed out that this rejection method is not optimal, since it ignores the information content of the rejected measurement. During the last 100 years, different weighting methods were developed that reduce the sensitivity of the least-squares estimator. This research led to the maximum likelihood method as one of the best linear estimators.

Noether (1967), Birnbaum/Laska (1967), Høyland (1968) show that the sample mean is very sensitive to grouping effects (e.g. tests or experiments done at different times at changed environmental conditions), and pairwise median estimators reduce the error covariance of the estimate up to 18% in the presence of gross errors.

Høyland (1968), Gastwirth/Rubin (1969), Parks (1967), and Jain (1975) show the sensitivity of estimates to dependencies and correlations between the errors.

2.4 CONFIDENCE MAPPING FOR ROBUST ESTIMATION IN STATIC SYSTEMS

In the following sections, the method of confidence mapping is introduced that makes the estimates less sensitive to the extreme tails of the data, with only a small sacrifice in optimality of the estimate. This method assumes that only statistics that belong to the open set, bounded by the confidence boundaries, follow the assumed error distribution function and errors outside the confidence boundaries belong to different and unknown distribution functions. Using this binary order statistic, the errors outside the confidence boundaries are mapped inside the boundaries by a confidence function and then treated as if they follow the assumed distribution of errors. For the estimation problem, the confidence boundaries have to be chosen such that the error inside the boundaries are sufficient statistics for the problem (generally around 2σ) and the confidence mapping function has to be chosen such that it does not violate the conditions for global optimality (convexity condition).

2.4.1 Effect of Deviations from the Assumed Probability Distribution

The mean of a random variable is

$$\mu = \int_{-\infty}^{\infty} x \; p(a,x) \; dx \; , \tag{2.3}$$

and its covariance is

$$\text{var } x = \int_{-\infty}^{\infty} x^2 \; p(a,x) dx, \tag{2.4}$$

where p(a,x) is the probability density function of the random variable x, and "a" is a parameter defining the density function from a particular class.

Deviations from the assumed distribution function (Prokhorov, 1956) occur when the random variables temporarily belong to a different distribution function. To determine the effect of deviations from the assumed distribution function we find the sensitivity of incremental contributions to mean and covariance due to distribution parameter changes.

The incremental contributions of particular values of the random variable, x, to mean and covariance are

$$\mu_x = x\, p(a,x) \tag{2.5}$$

$$v_x = x^2\, p(a,x) \tag{2.6}$$

and the sensitivities to distribution parameter changes are

$$\mu_{x_a} = x\, \frac{\partial p(a,x)}{\partial a} \tag{2.7}$$

$$v_{x_a} = x^2\, \frac{\partial p(a,x)}{\partial a} \tag{2.8}$$

The sensitivity of some of the common symmetric density functions are shown in Table 2.1. The highest sensitivity to deviations from the assumed distribution functions have normally distributed random variables. The sensitivity is small for random variables less than 2σ away from their mean, but it becomes large further away from the mean. The sensitivity of the mean of normally distributed random variables increases with the 3rd power of x, and the sensitivity of the variance with the 4th power. This explains the high sensitivity of the least squares method to deviations from the assumed probability distribution (Newcomb, 1886; Tukey, 1960; Huber, 1972).

Table 2.1

Sensitivities of Estimates Due to Changes in Distribution Parameters

DISTRIBUTION	CENSITY FUNCTION $p(a,x)$	SENSITIVITY CF MEAN $\frac{\partial}{\partial a}(\mu_x = xp(a,x))$	SENSITIVITY CF VARIANCE $\frac{\partial}{\partial a}(v_x = x^2p(a,x))$
NORMAL	$\frac{1}{\sqrt{2\pi}}\ \exp[-(x-\mu)^2/2a^2]$	$\left[\frac{(x-\mu)^2}{a^3} - \frac{1}{a}\right] xp(a,x)$	$\left[\frac{(x-\mu)^2}{a^3} - \frac{1}{a}\right] x^2p(a,x)$
EXPONENTIAL	$\frac{1}{2a}\exp\left[- \frac{\lvert x-\mu\rvert}{a}\right]$	$\left[\frac{\lvert x-\mu\rvert}{a^2} - \frac{1}{a}\right]xp(a,x)$	$\left[\frac{\lvert x-\mu\rvert}{a^2} - \frac{1}{a}\right]x^2p(a,x)$
BETA (symmetric)	$\frac{1}{B(a)}\ x^a(1-x)^a I(0,1)(x)$ $B(a) = \int_0^1 x^a(1-x)^a\ dx$	$\left[\ln(x(1-x)) - \frac{\partial B}{\partial a}\right] xp(a,x)$ $\frac{\partial B}{\partial a} = \int_0^1 \ln(x(1-x))x^a(1-x)^a\ dx$	$\left[\text{lu}(x(1-x)) - \frac{\partial B}{\partial a}\right] x^2p(a,x)$ $\frac{\partial B}{\partial a} = \int_0^1 \ln(x(1-x))x^a(1-x)^a\ dx$
UNIFORM	$\frac{1}{b-a}\ I(a,b)(x)$	$\pm xp(a,x)$	$\pm x^2p(a\ x)$

The sensitivity function of an exponential distribution function increases only with the 2nd power of x for the mean and 3rd power of x for the variance; it is therefore less sensitive to deviations from the nominal distribution function for large values of x or outliers. The sensitivity function of a uniform distribution increases only with 1st power of x for the mean and 2nd power of x for the variance. It is therefore least sensitive to distribution uncertainties.

The sensitivity functions show additionally that the mean of symmetrically distributed random variables is not changed by corruption from other symmetrically distributed random variables, since the sensitivity functions for the mean are of odd powers of x. The variance and the uncertainty in an estimate are of even power of x and are therefore affected by errors in distributions.

We also observe that the sensitivity increases with the third power of the inverse of the uncertainty for normally distributed random variables. This is of particular importance for the update of innovation processes, which are generally weighted with the inverse of the square of the estimate uncertainty. For the number of data $n \to \infty$ the estimate uncertainty becomes very small and therefore the ratio x/a becomes very large. Hence, the relative sensitivity to distribution uncertainties increases with the number of data points.

In summary, the more concentrated the assumed distribution is at its mean and the flatter the assumed tail, the more sensitive the estimates are to deviations from the assumed distribution.

In order to robustify an estimator, we have to reduce the sensitivity of the estimator to the tail distribution at the cost of a small increase in the variance of the estimate for nominal distribution.

2.4.2 Common Probability for Errors Outside and Inside the Confidence Boundaries

Errors within the measurement accuracy of a sensor system, e.g. errors within our confidence boundaries, can often be assumed normal or Poisson-distributed. Errors outside the confidence boundaries are generally due to failures, partial failures of components, and signal combinations not considered in the design of the measuring system. These errors are typically exponentially distributed or do not belong to any dense set of values. An estimator based on the exponential closure to this mixture of errors does not give sufficient weight to measurements with small errors. This estimator will be very robust to uncertainties in the distribution, but it will sacrifice on optimality, i.e. the uncertainty of the estimate will be unnecessarily large.

2.4.3 Rejection of Outliers

Legendre (1805) and Merrill (1972) treat errors outside some confidence boundaries as total failures of the measurement system and ignore the corresponding measurements for estimation purposes. Airy (1856), Ellis (1844), Fisher (1926), and Huber (1964) point out that environmental influences and partial failures often cause outliers which contain information in a degraded but not lost form. Rejecting these measurements eliminates the corresponding information contents and hence makes the estimator nonoptimal. Additionally, this method can lead to estimator instabilities for errors about the confidence boundaries.

De Laplace (1818) and Edgeworth (1886) showed that the median is the optimal estimator when errors follow no particular distribution and can be assumed to be uniformly distributed in a non-dense set of values. This is a non-parametric estimator that chooses the median random variable in an ordered set as the best estimate and rejects all other data. Therefore, its expected accuracy is directly proportional to the density of the measurements. Edgeworth (1886) showed also that the median is better than the least squares estimator for mixtures of Gaussian-distributed errors.

2.4.4 Confidence Mapping

Weighting the variables with some confidence measure that reduces the incremental influence of random variables from the tails of a distribution will robustify the estimator. This confidence mapping function (Salzwedel, Gupta, 1979) has to be chosen in such a way that it robustifies the estimator without destabilizing it for particular errors or groups of errors and leaves the estimator nearly optimal in the strict parametric sense.

Instead of using the estimate

$$E(x) = \int_{-\infty}^{\infty} x \, p(x) \, dx, \tag{2.9}$$

where $p(x)$ is the nominal probability density function, the estimate has now the form

$$E(x) = \int_{-\infty}^{\infty} x \, c(x) \, p(x) \, dx, \tag{2.10}$$

where $c(x)$ is a confidence mapping function of the form

$$c(x) = \frac{p_{robust}(x)}{p(x)} \quad . \tag{2.11}$$

The variance is then

$$var(x) = \int_{-\infty}^{\infty} x^2 \, c^2(x) p(x) \, dx \tag{2.12}$$

The rejection method of Section 2.4.3 can be seen as a confidence mapping function which maps the measurement into the a priori estimate and hence does not change the estimate.

Winsorizing maps the errors outside the confidence boundaries into the confidence boundaries. Huber's M-estimator uses a straight line continuation on the convex maximum likelihood function inside the confidence boundaries to reduce the destabilizing effect of outliers.

The estimation problem can be formulated in the form

$$\sum_i \rho(x_i, \hat{\theta}) = \min, \tag{2.13}$$

where $\rho(x,\theta)$ is some arbitrary function and $\hat{\theta}$ is the optimal estimate of the parameter θ.

If x_i belongs to a dense set, R_x, the problem can be stated in differentiated form

$$\sum_i \psi(x_i, \theta) = 0, \tag{2.14}$$

where

$$\psi(x,\theta) = \frac{\partial}{\partial\theta} \rho(x,\theta)$$

(Note: $\rho(x,\theta) = -\log f(x,\theta)$ gives an ordinary ML estimate.)

If θ is a location parameter, the problem becomes

$$\sum_i \rho(x_i - \hat{\theta}) = \min \tag{2.15}$$

or

$$\sum_i \psi(x_i - \hat{\theta}) = 0. \tag{2.16}$$

Equation (2.14) can be written

$$\sum_i w_i (x_i - \hat{\theta}) = 0; \tag{2.17}$$

with

$$w_i = \frac{\psi(x_i - \hat{\theta})}{x_i - \hat{\theta}}$$

we get the weighted mean

$$\hat{\theta} = \frac{\sum_i w_i x_i}{\sum_i w_i} \tag{2.18}$$

2.4.4.1 Huber's M-Estimator

Discussing Gauss's arithmetic mean (solution to $\sum_i (x_i - a)=0$), Ellis (1844) introduced a function $\psi(.)$ that gives different weight to measurements further away from the mean,

$$\Sigma\psi(x_i - a) = 0, \tag{2.19}$$

and brought up the question of choosing the function such as to obtain a stable estimator. Huber (1964) calls this estimator M-estimator (maximum likelihood estimator) and modifies the function ψ such that it corresponds to the ordinary inverse-log maximum likelihood function,

$$\rho(x,\theta) = - \log f(x,\theta) \tag{2.20}$$

and $\psi(x,\theta) = \frac{\partial}{\partial\theta}\rho(x,\theta)$, for random variables inside the confidence boundaries and a straight line continuation of $\rho(x,\theta)$ outside the confidence boundaries. This corresponds to a normal distribution function inside the confidence boundaries and an exponential distribution function outside. This likelihood function gives minimum weight to measures outside the confidence boundaries without violating the convexity condition for a global optimal estimate.

For normally distributed errors inside the confidence boundaries and exponentially errors outside, ρ, is of the form

$$\rho(x) = \begin{cases} \frac{x^2}{2} & \text{for } |x| \leq c \\ c|x| - \frac{c^2}{2} & \text{for } |x| > c \end{cases} \tag{2.21}$$

and

$$\psi(x) = \begin{cases} -c & \text{for } x < -c \\ x & \text{for } -c \leq x \leq c \\ c & \text{for } x > c \end{cases} \tag{2.22}$$

The estimator has thus the form

$$\sum_i w_i (x_i - \hat{\theta}) = 0 \tag{2.23}$$

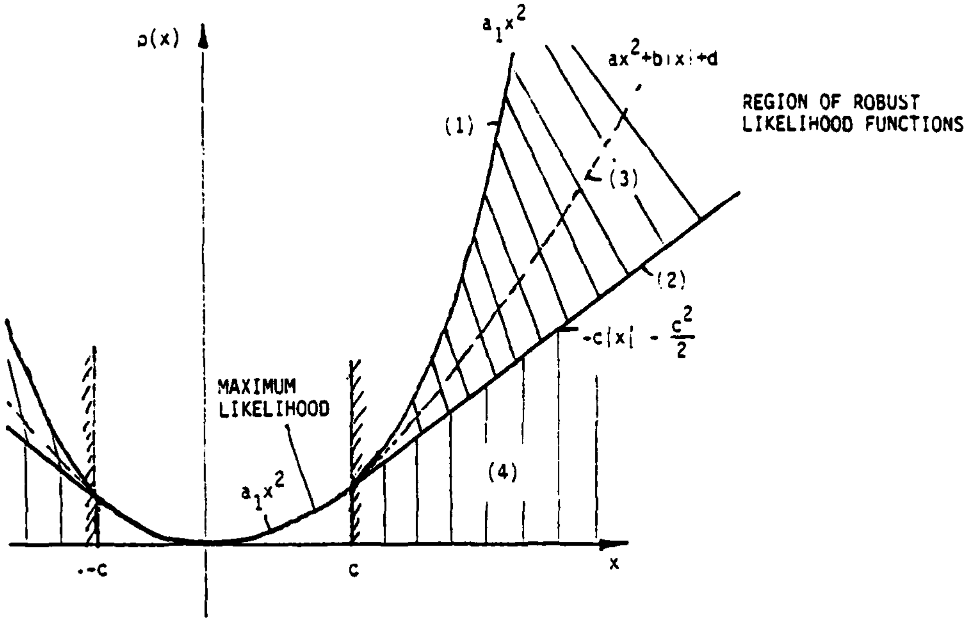
with weights w_i mapping errors outside the confidence boundaries on the confidence boundaries (Winsorizing), which is similar to the confidence mapping proposed by Newcomb (1886).

2.4.4.2 Robust Likelihood Functions in the Class of Linear Estimates

The condition for a linear estimator is that the likelihood function has no power higher than two. The condition for a unique optimum is that the likelihood function is convex. These two conditions require that the likelihood function has the form

$$\rho(x) = ax^2 + b|x| + d. \tag{2.24}$$

Any robust likelihood function must therefore be a linear combination of Eq. (2.24) (see Figure 2.1).



(1) Maximum likelihood function $\rho(x) = ax^2$

(2) Huber's robust likelihood function (2.10). This likelihood puts minimum weight on random variables outside the confidence boundaries without violating the convexity condition.

(3) Robust likelihood function.

(4) Region of robust likelihood functions that violate the convexity condition.

Figure 2.1 Robust Likelihood Functions

2.4.4.3 Robust Likelihood Estimates in a Class of Distribution Functions

In Section 2.4.4.1, we discussed Huber's robust M-estimator for a given distribution or density function. In the following the theory is extended for the case where the density function is unknown, but it is known to which class of density functions it belongs,

$$f \in F, \tag{2.25}$$

where F is a class of density functions, e.g. symmetric density functions, and f is a density function out of F, described by the parameter ϕ,

$$f = F(\phi). \tag{2.26}$$

The robustified likelihood function is then

$$\rho(x,\theta,\phi) = \begin{cases} \ell(f(x,\theta,\phi)) & \text{for } |x-\theta| \leq c \\ r(f(x,\theta,\phi)) & \text{for } |x-\theta| > c \end{cases} \tag{2.27}$$

where c is the confidence boundary, ℓ is a likelihood function for innovations $x-\hat{\theta}$ inside the confidence bounds (e.g. inverse-log likelihood function for purely exponential distributions f), and r is a robustified likelihood function outside the confidence boundaries, that reduces the sensitivity of the estimator due to outliers and deviations of random values, x, from the assumed distribution.

The optimal estimate θ of the parameter $\hat{\theta}$ is the solution of

$$\sum_i \rho(x_i,\hat{\theta},\hat{\phi}) = \min \tag{2.28}$$

If $\rho(x,\hat{\theta},\hat{\phi})$ is convex and the derivative

$$\psi(x,\theta,\phi) = \frac{\partial}{\partial\theta}\rho(x,\theta,\phi) + \frac{\partial}{\partial\phi}\rho(x,\theta,\phi)$$

is continuous and bounded, Eq. (2.28) reduces to

$$\sum_i \psi(x_i,\hat{\theta},\hat{\phi}) = 0, \tag{2.29}$$

and in the case of a location parameter to

$$\sum_i \psi(x_i-\hat{\theta},\phi) = 0 \tag{2.30}$$

2.4.5 Robustness of Estimators for Finite Density Functions

Finite density functions have nonzero values only for random values $x_{min} \leq x \leq x_{max}$. Therefore, maximum likelihood functions of estimators for random variables with finite density functions have the form

$$\rho(x,\theta) = \begin{cases} \text{zero} & \text{for } x < x_{min} \\ \ell(x,\theta) & \text{for } x_{min} \leq x \leq x_{max} \\ \text{zero} & \text{for } x > x_{max} \end{cases} \tag{2.31}$$

Hence, maximum likelihood estimators for finite random variables have natural confidence boundaries, and values outside the confidence boundaries are considered outliers and rejected for the estimation problem. The robustness of maximum likelihood estimators for finite random variables is inversely proportional to the difference between the confidence boundaries.

2.5 APPLICATION TO DYNAMIC SYSTEMS

The concepts presented above for static systems may be extended for dynamic systems. This section shows the difficulties in this extension and how some of the problems are resolved.

Consider a dynamic system in discrete form with state x, controls u, outputs y, and parameters θ.

$$x_{k+1} = f(x_k,u_k,\theta) + w_k \quad k=0,1,2...N-1 \tag{2.32}$$

$$y_k = h(x_k,u_k,\theta) + v_k \quad k=1,2,...,N \tag{2.33}$$

w_k and v_k are process and measurement noise sources. The standard procedures for the estimation of parameters θ are based on the assumption that w_k and v_k are white with Gaussian densities.

In general, this problem can be solved as an estimation problem along the time-coordinate in n+1-dimensional space. The n-dimensional state x of the dynamic system is known with some uncertainty Π_o at time $t=t_o$. Using the m-dimensional parameter state of the sytem the state is predicted for the time $t=t_1=t_o+\Delta t_o$. Using $\hat{\theta}$, the uncertainty Π_o is mapped into the n-dimensional plane at $t=t_1$, $\Pi_o \rightarrow \Pi_1$. Because of parameter errors and disturbances on the system, the uncertainty increases to $\Pi_1^2 = \Pi_1 \cap \Pi_p$. A measure y_1 is taken and an innovation $\nu_1 = y_1 - h_1(\hat{x}_1,\hat{\theta})$ is formed. The problem we face now is to decide where the confidence region lies. It is better to expand the confidence region about the predicted state or the measurement, or should we expand it about the updated state, formed using the assumption that the measurement as well as the predicted state are included in the assumed probability space? As long as the confidence region Π^c of the measurement is included in the state uncertainty, $\Pi^c \subset \Pi^2$, it cannot be detected whether a particular measurement falls within its tail-statistics if it is also included in Π^2.

We shall first look at the case where the process noise $\Pi^p \subset \Pi^c$. Then, after a sufficient number of measurements $\Pi^2 \subset \Pi^c$, and the confidence region can be expanded about the predicted state $\hat{x}$ and confidence mapping can be applied to the innovation sequence, ν (Hampel, 1971; Papantoni-Kazakos/Gray, 1978).

2.6 NO PROCESS NOISE

When there is no process noise, estimators which are robust for deviations from the nominal measurement noise probability density may be designed. Following along the lines of the static estimation, a cost functional is defined in the following manner,

$$J = \sum_{k=1}^{N} [e_k^T R^{-1}(e_k) e_k + \log|R(e_k)|] \tag{2.34}$$

The weighting R is a function of the error itself. If the errors in the measurements are known to be uncorrelated, the matrix R is diagonal. The functional forms for the diagonal terms of R must be such that the dimensions of the particular measurement do not affect the confidence bound. Therefore

$$\xi_k(i) = \frac{e_k(i)}{\alpha_i} \tag{2.35}$$

$$R_{ii}(e_k) = \alpha_i^2 \rho(\xi_k(i)) \tag{2.36}$$

where $e_k(i)$ is the ith component of the error vector at time k. The function $\rho(\cdot)$ is the same for each measurement. To estimate α_i, we differentiate Eq. (4.3) with respect to α_i and set the result to zero (assuming $\rho(\xi_k(i))$ is differentiable everywhere).

$$\frac{\partial J}{\partial \alpha_i} = \sum_{k=1}^{N} \left[- \frac{e_k^2(i)}{[\alpha_i^2 \rho(\xi_k(i))]^2} + \frac{1}{[\alpha_i^2 \rho(\xi_k(i))]} \right] \cdot \left[\alpha_i^2 \frac{\partial \rho(\xi_k(i))}{\partial \alpha_i} + 2\alpha_i \rho(\xi_k(i)) \right] = 0 \tag{2.37}$$

and

$$\frac{\partial\rho(\xi_k(i))}{\partial\alpha_i} = \frac{\partial\rho(\xi_k(i))}{\partial\xi_k(i)} \cdot \left[- \frac{\ell_k(i)}{\alpha_i^2} \right]$$

$$= \frac{\partial\rho(\xi_k(i))}{\partial\xi_k(i)} \cdot \left[- \frac{\xi_k(i)}{\alpha_i} \right] \qquad (2.38)$$

Therefore

$$\alpha_i^2 = \frac{\sum_{k=1}^{N} w_k(i) \frac{e_k^2(i)}{\rho(\xi_k(i))}}{\sum_{k=1}^{N} w_k(i)} \qquad (2.38)$$

with

$$w_k(i) = 2 - \frac{\xi_k(i)}{\rho(\xi_k(i))} \frac{\partial\rho(\xi_k(i))}{\partial\xi_k(i)} .$$

If ρ is a mild function of the variable $\xi_k(i)$, the above equation may be simplified to

$$\alpha_i^2 = \frac{1}{N} \sum_{k=1}^{N} \frac{\ell_k^2(i)}{\rho(\xi_k(i))} . \qquad (2.40)$$

The estimation of parameters based on the likelihood function of Eq. (2.34) must therefore be done in two steps.

(1) Select an α_i and perform a Newton-Raphson optimization to estimate the system parameters. In this step, derivatives of $\rho(\xi_k(i))$ with respect to system parameters may be ignored.

(2) Use the above equations to estimate α_i. Repeat the procedure till convergence occurs.

This technique may be modified for the case of gaussian process noise and nongaussian measurement noise.

2.7 PROCESS NOISE

To develop procedures for Non-Gaussian process noise, we assume that its distribution is generally normal with some contamination from another normally distributed noise of higher standard deviation. Suppose at any point in time, the process noise is a realization of the higher variance random variable. The state at the next point will be highly perturbed from its expected value. This perturbation may reduce in the following step, if the system is stable. Nevertheless, its effect will be felt in a number of subsequent steps. This is very different from the case of the non Gaussian measurement noise where the effect of each deviation is felt only in the particular step.

The previous discussion points to two aspects of problems with nongaussian noise. The nongaussian effect (particularly the case when one basic distribution is contaminated by another distribution with higher variance) at one sample point lasts over many future points. This complicates the problem. However, because of this reason, it appears that more information is available to isolate nongaussian behavior.

The critical problem in the development of an algorithm is to ensure that the nongaussian effects at one point do not influence the estimates of noise sources at other points. The most direct procedure to achieve this objective is to identify the system parameters as well as the process noise variable w without making any a priori assumption about w. In other words, w is assumed to be a completely unknown input signal with unspecified characteristics. With no assumptions on w, it is possible to find its estimate only when the number of measurements exceeds the number of process noise sources.

Let $\hat{w}_k$ $(k = 1,2,3,\ldots N)$ be the estimate of w based on no a priori information. Using this estimate, it is possible to select the values of k for which w_k may be assumed to come from a Gaussian distribution. A Gaussian covariance may be specified for process noise at those points while the other points are assumed completely unknown. This procedure is repeated till convergence occurs.

The difficulty with using this procedure is its integration with the parameter estimation algorithm. It appears that the procedure for specifying $\hat{w}_k$ has to be reinitialized following each parameter iteration. The computation time requirements may be unacceptable. It may be sufficient, however, to update $\hat{w}_k$ once after several iterations.

2.8 EXAMPLE OF ROBUST LIKELIHOOD ESTIMATION

Many problems have been solved by the use of robust maximum likelihood methods. This section presents an example to demonstrate the improvement in estimation accuracy which may be achieved by the application of these methods. Monte Carlo methods are used to demonstrate a quantitative reduction in variance.

Consider the following nonlinear system.

$$\begin{aligned} x_1 &= \theta_{11}\, x_1 + \theta_{12}\, x_1^2 + \theta_{13}\, x_2 \\ &\quad + \theta_{14}\, x_2^2 + u_1 \\ x_2 &= \theta_{21}\, x_1 + \theta_{22}\, x_1^2 + \theta_{23}\, x_2 \\ &\quad + \theta_{24}\, x_2^2 + u_2 \end{aligned} \tag{2.42}$$

with parameters

$$\theta = \begin{bmatrix} -1.0 & .01 & .2 & -.02 \\ -.15 & .03 & -1.0 & .015 \end{bmatrix} \tag{2.43}$$

and forcing function

$$u = \begin{bmatrix} 0 \\ t\ e^{-.1t} \end{bmatrix} \tag{2.44}$$

was observed by measurements

$$y = x + n.$$

The errors in the measurements are a mixture of normally distributed random variables

$$n_1 = N\left(0, \begin{bmatrix} .05 & 0 \\ 0 & 1.0 \end{bmatrix}\right) \tag{2.46}$$

and

$$n_2 = N\left(0, \begin{bmatrix} .5 & 0 \\ 0 & 10.0 \end{bmatrix}\right) \tag{2.47}$$

$$n = \begin{cases} n_1 & 90\% \text{ of the time} \\ n_2 & 10\% \text{ of the time} \end{cases} \qquad (2.48)$$

Figure 2.2 shows maximum likelihood and robust likelihood estimates of the parameters θ_{11} and θ_{23} respectively, for different $n^i \varepsilon \mathcal{N}$, the space of all random variables, defined by Eq. (2.48). The variances of the parameter estimates and the output errors are

	MAXIMUM LIKELIHOOD	ROBUST LIKELIHOOD	MAXIMUM LIKELIHOOD ROBUST LIKELIHOOD
$\sigma_{\theta_{11}}$	.095	.033	2.9
$\sigma_{\theta_{23}}$	.201	.050	3.7
σ_{y_1}	.02758	.02804	.98
σ_{y_2}	9.205	9.205	.99

[NUMBER OF DATA POINTS N=50]

The robust likelihood estimation gave parameter estimates for this example that are 2.9 and 3.7 times better than the parameter estimates of the maximum likelihood estimator with an average increase in output error of less than 2% and a maximum increase of output error of less than 5%.

When the number of parameters is large or if some parameters are marginally identifiable, further improvements in estimation accuracy may be achieved by the use of robust procedures.

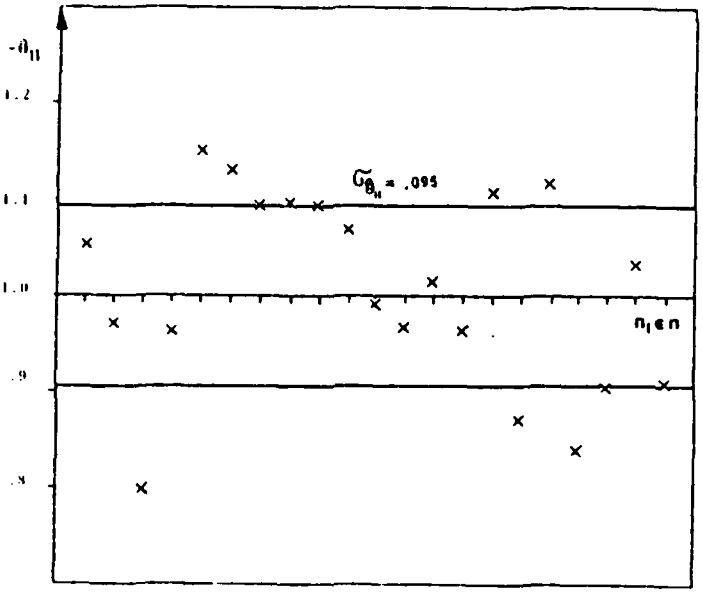


Figure 2.2a Maximum Likelihood Estimate of Parameter θ_{11}

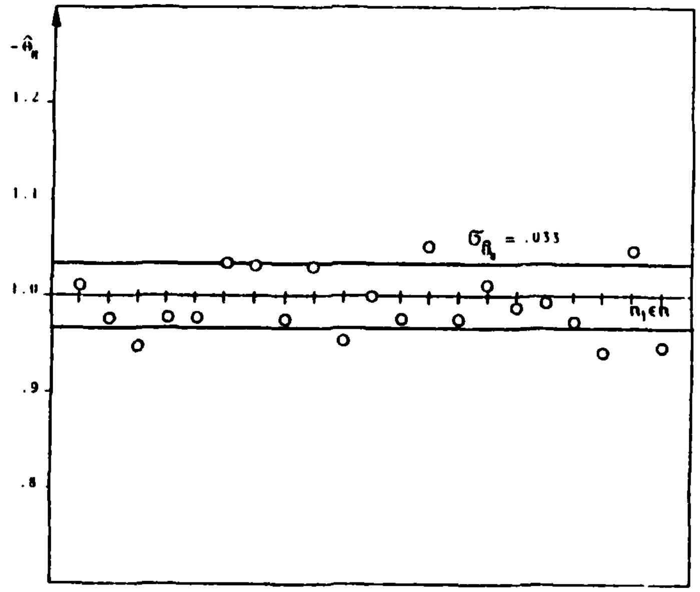


Figure 2.2b Robust Estimate of Parameter θ_{11}

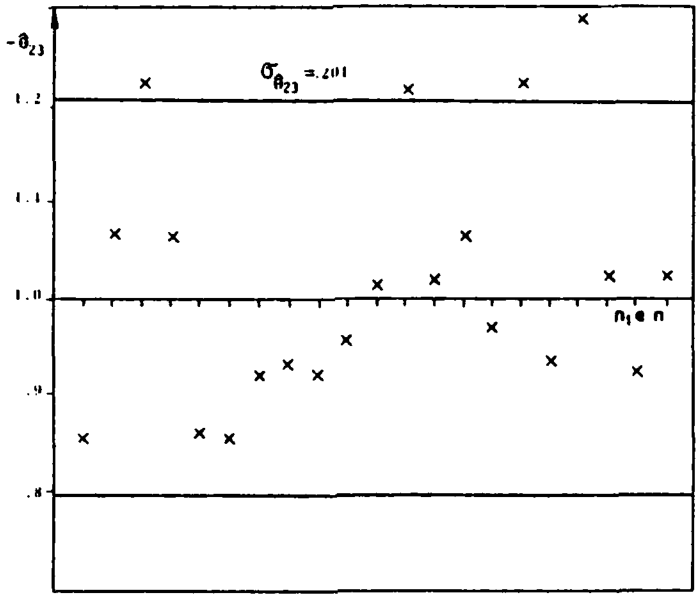


Figure 2.2c Maximum Likelihood Estimate of Parameter θ_{23}

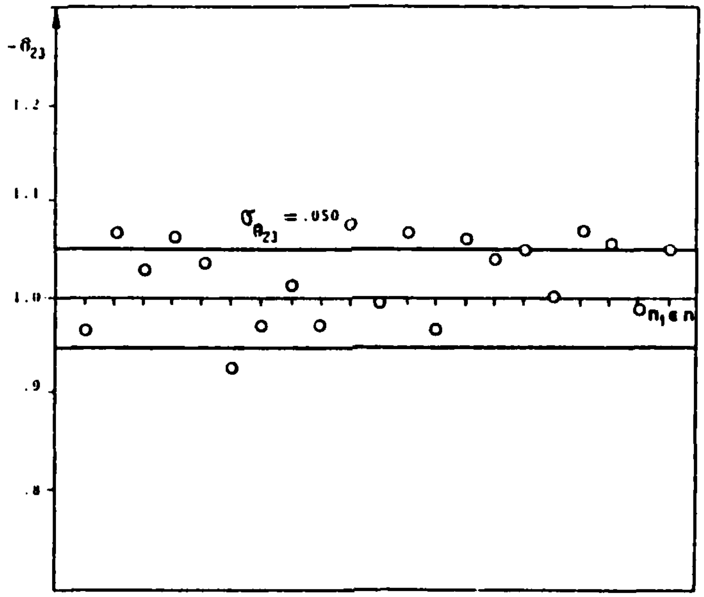


Figure 2.2d Robust Estimate of Parameter θ_{23}

2.9 CONCLUSIONS

Accuracy of estimates resulting from experimental data may be significantly enhanced by applying robust techniques. These techniques require modification to the least squares type of performance index in general and to the maximum likelihood method in particular. Huber's work [15-17] provides good background for the development of robust methods for dynamic systems.

Though this section has been concerned with parameter estimation, similar methods apply for state estimation also. Application of these methods could provide significant improvement in dynamic state estimation when Kalman filters are used.

III. PARAMETER ESTIMATION WITH NON-WHITE ERRORS

3.1 INTRODUCTION

Measurement errors are generally correlated due to instrument dynamics, finite moments of inertia, dynamics and environmental influences. The environmental influences disturb the system and the measurement instruments at the same time or with some time-delay and hence correlate the system noise and the measurement error.

The nonwhiteness and the cross-correlation between system and measurement errors correlate the innovation process in the Kalman filter unless these effects are compensated for. It is shown that estimates will have increased covariances but will not be biased. Techniques for including the effects of known measurement error correlations in state estimation problems have been considered by several previous authors (Kalman, 1963; Henrikson, 1968; Jazwinski, 1970). This chapter describes state as well as parameter estimation methods for non-white noise.

3.2 PROBLEM FORMULATION

In non-white noise problems, the spectrum of the process and measurement noise sources is often not known. Without a priori knowledge of the noise power spectrum, one of the following is required:

(1) Estimate the correlation and adapt the state and parameter estimator to the determined correlation.

(2) Sacrifice some of the sensitivity of the estimator to make it insensitive (robust) against a set of possible correlations of the worst case correlation.

In the first approach, an additional spectrum estimation step is required. Once the spectrum is known, procedures for the known spectrum can be used for state and parameter estimator.

The second approach leads to a worst case estimator. A certain efficiency is lost if the noise is white but the estimators are robust with non-white noise, particularly when the noise has a spectrum similar to that of the signal.

3.3 ESTIMATORS FOR KNOWN CORRELATIONS BETWEEN ERRORS

Let the discrete linear system (A non-linear system can be approximated by piece-wise linear systems) be

$$x_{k+1} = \phi_k x_k + \Gamma_k w_k , \quad k = 0,1,\ldots,N-1, \qquad (3.1)$$

with measurement outputs

$$y_k = H_k x_k + v_k , \quad k = 1,2,\ldots,N, \qquad (3.2)$$

where

$x_k \varepsilon R^n$ is the system state,

$y_k \varepsilon R^m$ is the measurement of the output $H_k x_k$,

$w_k \varepsilon R^p$ noise sequence of the system

$v_k \varepsilon R^m$ noise sequence of the measurements

$\bar{x}_k = E[x_k]$

$E[w_k] = E[v_{k+1}] = 0 \quad k = 0,1,\ldots,N-1$

Define the correlation matrices

$$\begin{aligned}
R_{xx}(k,j) &\hat{=} E[(x_k-\bar{x}_k)(x_j-\bar{x}_j)^T], \\
R_{ww}(k,j) &\hat{=} E[w_k w_j^T], \\
R_{vv}(k,j) &\hat{=} E[v_k v_j^T], \qquad (3.3) \\
R_{wv}(k,j) &\hat{=} E[w_k v_j^T], \\
P_{k+1}^k &\hat{=} E[(x_{k+1}-\hat{x}^k{}_{k+1})(x_{k+1}-\hat{x}^k{}_{k+1})^T | Y_k],
\end{aligned}$$

where

$$\hat{x}^k_{k+1} \mathrel{\hat{=}} E[x_{k+1}|Y_k].$$

For the white noise ease with a priori known covariance,

$$E[w_k w_j{}^T] = R_{ww}(k,k)\delta_{kj},$$

$$E[v_k v_j{}^T] = R_{vv}(k,k)\delta_{kj},$$

$$E[w_k v_j{}^T] = 0,$$

the minimum variance filter is between measurements,

$$\hat{x}^k_{k+1} = \phi_k \; \hat{x}^k_k \tag{3.4}$$

$$P^k_{k+1} = E[(\phi_k(x_k - \hat{x}^k_k) + \Gamma_k w_k)(\phi_k(w_k - \hat{x}^k_k) + \Gamma_k w_k)^T|Y_k]$$

$$P^k_{k+1} = \phi_k P^k_k \phi^T_k + \Gamma_k R^k_{ww} \;)\Gamma_k{}^T, \tag{3.5}$$

and at observations

$$\hat{x}^k_k = \hat{x}^{k-1}_k + K_k(y_k - H_k \; \hat{x}^{k-1}_k), \tag{3.6}$$

$$P^k_k = P^{k-1}_k - K_k \; H_k \; P^{k-1}_k \;, \tag{3.7}$$

where

$$K_k = P^{k-1}_k H^T{}_k [H_k P^{k-1}_k H_k{}^T + R_{vv}{}^k]^{-1} \tag{3.8}$$

is the Kalman gain.

3.3.1 Correlated State and Measurement Noise with A Priori Known Covariance

For the a priori known covariance, $E[w_k v_j{}^T] = R_{wv}(k,k)\delta_{kj} \neq 0$, between state and measurement noise an optimal estimator can be assigned by the method of Kalman (1963), and Jazwinski (1970).

Let $U_k = \{u_1,...,u_k\}$ be an orthonormal basis for the measurement Y_k,s.t. $E[u_i u_k^T] = I\ \delta_{ik}$. Since the best estimate $\hat{x}^k_{k+1}$ of x_{k+1} is the orthogonal projection of x_{k+1} into Y , it can be described by

$$\hat{x}^k_{k+1} = \sum_{i=1}^{k} E[x_{k+1}u_i^T]u_i \quad , \qquad (3.8)$$

and with (3.1)

$$\hat{x}^k_{k+1} = \sum_{i=1}^{k} E[(\phi_k \cdot x_k + \Gamma_k w_k)u_i^T]u_i + E[x_{v+1}u_k^T]u_k . \qquad (3.9)$$

Since w_k is not sequentially correlated, it is independent of Y_{k-1}, and

$$\hat{x}^k_{k+1} = \phi_k\ \hat{x}^{k-1}_k + E[x_{k+1}u_k^T]u_k . \qquad (3.10)$$

Because the innovation $\nu_k = y_k - H_k\hat{x}^{k-1}_k$ is orthogonal to Y_{k-1} but included in Y_k,

$$E[x_{k+1}u_k^T]u_k = K^*{}_k[y_k - H_k\hat{x}^{k-1}_k] . \qquad (3.11)$$

The error in the estimate, $\tilde{x}^k_{k+1} \hat{=} x_{k+1} - \hat{x}^k_{k+1}$, is independent of the measurement y_k, hence $E[\tilde{x}^k_{k+1}\ y_k^T] = 0$. This gives,

$$K^*{}_k = [\phi_k P^{k-1}_k H_k^T + \Gamma_k R_{wv_k}][H_k P^{k-1}_k H_k^T + R_{vv_k}]^{-1} \qquad (3.12)$$

with

$$P^k_{k+1} \hat{=} E[\tilde{x}^k_{k+1}\ \tilde{x}^k_{k-1}{}^T] = \phi_k P^{k-1}_k \phi_k^T + \Gamma_k R_{ww}\Gamma_k^T \qquad (3.13)$$

$$- K^*{}_k[H_k P^{k-1}_k \phi_k^T + R_{wv_k}{}^T \Gamma_k{}^T] .$$

The filter for correlated state and measurement noise is then,

<u>Measurement update</u>:

$$\hat{x}^k_k = \hat{x}^{k-1}_k + K_k(y_k - H_k\hat{x}^{k-1}_k) \qquad (3.14)$$

$$P^k_k = P^{k-1}_k - K_k H_k P^{k-1}_k \qquad (3.15)$$

$$K_k = P^{k-1}_k H_k^T[H_k P^{k-1}_k H_k^T + R_{w_k}]^{-1} \qquad (3.16)$$

Time update:

$$\hat{x}^k_{k+1} = \phi_k \hat{x}^k_k + \Gamma_k R_{wv_k} [H_k P^{k-1}_k H_k{}^T + R_{vv_k}]^{-1} (y_k - H_k \hat{x}^{k-1}_k) \tag{3.17}$$

$$P^k_{k+1} = \phi_k P_k{}^k \phi_k{}^T + \Gamma_k R_{ww_k} \Gamma_k{}^T - \Gamma_k R_{ww_k} [H_k P^{k-1}_k H_k{}^T + R_{vv_k}]^{-1} \tag{3.18}$$

$$- R_{wv}{}^T \Gamma_k{}^T - \phi_k K_k R_{wv}{}^T \Gamma_k{}^T - \Gamma_k R_{wv} K_k{}^T \phi_k{}^T .$$

3.3.2 Sequentially Correlated Measurement Noise with A Priori Known Covariance

Let the measurement noise be correlated through the Markov sequence

$$v_{k+1} = \Psi_k v_k + u_k \tag{3.19}$$

with driving noise $u_k \sim N(0, R_{uu_k})$ uncorrelated to the state noise w_k, $E[u_k w_k{}^T] = 0$. Kalman (1960) solves this problem by augmenting the Markov sequence (3.19) to the state equation (3.1):

$$x^a \hat{=} \begin{bmatrix} x \\ v \end{bmatrix}, \qquad \phi^a \hat{=} \left[\begin{array}{c|c} \phi & 0 \\ \hline 0 & \Psi \end{array}\right], \ \Gamma^a \hat{=} \left[\begin{array}{c|c} \Gamma & 0 \\ \hline 0 & I \end{array}\right] \ w^a \hat{=} \begin{bmatrix} w \\ u \end{bmatrix},$$

$$H^a \hat{=} [H \quad I], \qquad R^a_{ww} \hat{=} \begin{bmatrix} R_{ww} & 0 \\ 0 & R_{uu} \end{bmatrix},$$

$$x^a_{k+1} = \phi^a{}_k \, x^a{}_k + \Gamma_k{}^a \, w^a{}_k , \tag{3.20}$$

with noiseless measurements

$$y^a{}_k = H^a{}_k \, x^a{}_k . \tag{3.21}$$

In the formulation (3.20), (3.21), the error covariance of the estimate becomes singular. To overcome this problem, Bryson/Henrikson (1968) and Bryson/Ho (1969) used the difference of successive measurements, which has additive white noise.

$$\zeta_k \hat{=} y_k - \Psi_{k-1} y_{k-1} \tag{3.22}$$

$$\zeta_k = H^*_{k-1} x_{k-1} + u^*_{k-1} \tag{3.23}$$

with

$$H^*_{k-1} \hat{=} H_k \, \phi_{k-1} - \Psi_{k-1} \, H_{k-1}$$

$$u^*_{k-1} \hat{=} H_k \, \Gamma_{k-1} \, w_{k-1} + u_{k-1}$$

and

$$R^*_{uu_{k-1}} \hat{=} E[u^*_{k-1} u^*_{k-1}{}^T] = H_k \Gamma_{k-1} R_{ww_{k-1}} \Gamma_{k-1}{}^T H_k{}^T + R_{uu_{k-1}} .$$

The system is now

$$\begin{aligned} x_{k+1} &= \phi_k \, x_k + \Gamma_k \, w_k \\ \zeta_{k+1} &= H^*_k \, x_k + H_{k+1} \Gamma_k \, w_k + u_k , \qquad k \geq 1 \\ H^*{}_k &= H_{k+1} \phi_k - \Psi_k \, H_k \end{aligned} \tag{3.24}$$

with $v_o \sim N(0, R_{vv_o})$ and $E[x_o v_o{}^T] = 0$, we get from the augmented equations (2.20, 2.21)

$$\begin{aligned} \hat{x}^1_1 &= \hat{x}_o + P_o \, H_1{}^T [H_1 P_o H_1{}^T + R_{vv_o}]^{-1} (y_1 - H_1 \hat{x}_o) \\ P^1_1 &= P_o - P_o H_1{}^T [H_1 P_o H_1{}^T + R_{vv_o}]^{-1} H_1 P_o . \end{aligned} \tag{3.25}$$

The filter for system (2.34) is found by the method of Section (3.3.1),

$$\begin{aligned} \hat{x}^{k+1}_{k+1} = \; & \phi_k \hat{x}^k_k + [\phi_k P_k{}^k H^*{}_k{}^T + \Gamma_k R_{wu^*_k}] \\ & \cdot [H^*_k \, P^k{}_k H^*{}_k{}^T + R^*_{uu}]^{-1} [\zeta_{k+1} - H^*_k \, \hat{x}^k_k] \end{aligned} \tag{3.26}$$

$$\begin{aligned} P^{k+1}_{k+1} = \; & \phi_k P^k_k \phi_k{}^T + \Gamma_k R_{ww_k} \, \Gamma_k{}^T \\ & - [\phi_k P_k{}^k H^{*T}_k + \Gamma_k R_{wu^*_k}] [H^*_k P_k{}^k H^{*T}_k + R^*_{uu}]^{-1} [H^*_k P_k{}^k \phi_k{}^T + \\ & R_{ww}{}^T \Gamma_k{}^T] \end{aligned} \tag{3.27}$$

with

$$R_{wu*} = R_{ww_k} \Gamma_k{}^T H_{k+1}{}^T \qquad (3.28)$$

$$R^*_{uu} = H_{k+1} \Gamma_k R_{ww_k} \Gamma_k{}^T H_{k+1}{}^T + R_{uu_k}$$

The solution for time correlated measurement noise of continuous systems is shown by Bryson/Ho (1969) p. 405ff, Mehra/Bryson (1968), Bryson/Johansen (1965)

3.4 CORRELATED ERRORS WITH UNKNOWN COVARIANCE

3.4.1 Impact of Correlations

Comparing the optimal estimator for white noise (3.3 - 3.7) with the estimators for correlated measurement and state noise (3.14 - 3.18) and sequentially correlated measurement noise, we observe the following:

(1) Contrary to the white noise case, the time update of the optimal filter for the correlated noise requires feedback of the innovation sequence, $y - H\hat{x}$.

(2) If the measurement noise couples with the state equations, the Kalman gain for the optimal filter is different than the gain for uncorrelated errors.

To further investigate the impact of correlated noise on the estimation problem, the worst type of correlation is determined in Section 3.4.2, and its effect on a parameter estimation that assumes uncorrelated noise is investigated.

3.4.2 Worst Type Correlation Function

We first consider a system without process noise, but with additive measurement noise. Let the system be defined by

$$\dot{x} = f\ (x, \theta, u, t) \tag{3.29}$$

$$y = h\ (x, \theta, u, t)\ +\ \nu, \tag{3.30}$$

where x: state space
θ: parameters
u: control
ν: additive measurement noise with covariance R.

To obtain the maximum likelihood estimate, we minimize

$$J = \int_t (y-h)^T R^{-1} (y-h) dt,\ t_{min} \leq t \leq t_{max} \tag{3.31}$$

The worst case noise maximizes estimation error and hence minimizes the information matrix. Since functional characteristic of non-white noise is best described by its frequency distribution, we transform our estimation problem into the frequency domain. Assuming all functions have Fourier transforms, the cost function in the frequency domain is,

$$J_f = \int_{-\infty}^{\infty} (y-h)^* S_{vv}^{-1} \quad (y-h)\ d\omega\ . \tag{3.32}$$

The information matrix for the parameters θ is then

$$M_f = \int_{\omega_{min}}^{\omega_{max}} \frac{\partial h^*}{\partial \theta} S_{vv}^{-1} \frac{\partial h}{\partial \theta} d\omega\ , \tag{3.33}$$

where

$$\text{cov}\ \hat{\theta} = M^{-1}.$$

We now find a frequency distribution of the noise spectrum such that the information matrix, M , is a minimum and the covariance of the noise is a constant.

$$\min_{S_{vv}(\omega)} M \; \Big| \; \int_{\omega_{max}}^{\omega_{max}} S_{vv}(\omega)\, d\omega = \text{constant.} \qquad (3.34)$$

Adjoining the constraint and the information matrix gives

$$M = \int_{\omega_{max}}^{\omega_{max}} \left\{ \frac{\partial h^*}{\partial \theta} S_{vv}^{-1}(\omega) \frac{\partial h}{\partial \theta} + \Lambda S_{vv} \right\} d\omega\text{-const.} \qquad (3.35)$$

A good measure of matrix M is its trace,

$$\text{tr } M = \int_{\omega} \text{tr}\{ \frac{\partial h^*}{\partial \theta} S_{vv}(\omega)^{-1} \frac{\partial h}{\partial \theta} + \Lambda\, S_{vv}(\omega)\}\, d\omega - c \qquad (3.36)$$

tr M is a minimum if

$$\frac{\partial \text{tr} M}{\partial S_{vv}} = \int_{\omega} \{-(S_{vv}^{-1} \frac{\partial h}{\partial \theta} \frac{\partial h^*}{\partial \theta} S_{vv}^{-*})^* + \Lambda^*\}\, d\omega = 0 \qquad (3.37)$$

For the integral to be zero, the expression under the integral has to be zero. We get, therefore,

$$S_{vv}(\omega)\, \Lambda\, S_{vv}^*(\omega) = \frac{\partial h(\omega)}{\partial \theta} \frac{\partial h^*(\omega)}{\partial \theta} \;\hat{=}\!: H(\omega) \;, \qquad (3.38)$$

$$\dim \theta \geq \dim v \;,$$

$$\int_{\omega} S_{vv}(\omega)\, d\omega = \text{const} = S. \qquad (3.39)$$

Since both $S_{vv}(\omega)$ and $H(\omega)$ are symmetric, Λ also has to be symmetric, and we can decompose the equation

$$(S_{vv}\Lambda^{1/2})\,(S_{vv}\Lambda^{1/2})^* = H = H^{1/2}\, H^{*1/2}; \qquad (3.40)$$

hence,

$$S_{vv}(\omega) = \Lambda^{-1/2}\, H^{1/2}\;. \qquad (3.41)$$

Using the constraint

$$\Lambda^{-\frac{1}{2}} \int_{\omega} H^{\frac{1}{2}}(\omega)\ d\omega = S \quad , \tag{3.42}$$

the language multiplier is

$$\Lambda^{\frac{1}{2}} = \{ \int_{\omega} H^{\frac{1}{2}}(\omega)\ d\omega \}\ S^{-1} \quad . \tag{3.43}$$

Then the spectral noise distribution for the extremum of the information matrix is

$$S_{VV}(\omega) = [\{ \int_{\omega} H^{\frac{1}{2}}(\omega) d\omega \} S^{-1}]^{-1}\ H^{\frac{1}{2}}(\omega) \tag{3.44}$$

$$S_{VV}(\omega) = S\{ \int_{\omega} H^{\frac{1}{2}}(\omega) d\omega \}^{-1} H^{\frac{1}{2}}(\omega) \quad . \tag{3.45}$$

Writing the equation in the form

$$\{ \int_{\omega} S_{VV}(\omega) d\omega \}^{-1} S_{VV}(\omega) = \{ \int_{\omega} H^{\frac{1}{2}}(\omega) d\omega \}^{-1} H^{\frac{1}{2}}(\omega) \tag{3.46}$$

shows that the information matrix M is a minimum if the distribution of $S_{VV}(\omega)$ is the same as the distribution of $H^{\frac{1}{2}}(\omega)$. For an estimator of this type, the estimates follow the weighted noise in a maximum fasion, as observed in many estimation problems. The residual errors of the estimates will then be minimized and give the illusion of a good parameter estimate; even so, the parameters just follow the noise. Clearly, estimation errors may be significantly increased if the noise is the same frequency range as the signal.

3.4.3 Effect of Worst Case Noise Distribution on Estimation Error

The information matrix is

$$\mathrm{tr}\ M_f = \int_{\omega_{min}}^{\omega_{max}} \frac{\partial h}{\partial \theta}^{*} S_{vv}^{-1} \frac{\partial h}{\partial \theta}\ d\omega \quad , \tag{3.47}$$

and the trace of the information matrix is

$$\left.\begin{aligned} \text{tr } M_f &= \int_{\omega_{min}}^{\omega_{max}} \text{tr}\{\frac{\partial h}{\partial \theta}^* S_{vv}{}^{-1}\frac{\partial h}{\partial \theta}\}d\omega \\ &= \int_{\omega_{min}}^{\omega_{max}} \text{tr}\{\frac{\partial h}{\partial \theta} \ \frac{\partial h}{\partial \theta}^* S_{vv}{}^{-1}\}d\omega \\ \text{tr } M_f &= \int_{\omega_{min}}^{\omega_{max}} \text{tr}\{H(\omega)S_{vv}{}^{-1}(\omega)\} \ d\omega \ . \end{aligned}\right\} \qquad (3.48)$$

The worst case noise of the optimal filter is

$$S_{vv}(\omega) = S(\overline{H}^{\frac{1}{2}})^{-1}H^{\frac{1}{2}}(\omega)$$

$$S_{vv}{}^{-1}(\omega) = (H^{*\frac{1}{2}}(\omega))^{-1}\overline{H}^{\frac{1}{2}}S^{-1} \quad d$$

$$\text{tr } M_f = \int_{\omega_{min}}^{\omega_{max}} \text{tr}\{H^{\frac{1}{2}}(\omega)\overline{H}^{\frac{1}{2}}S^{-1}\} \ d\omega$$

$$\text{tr } M_f = \text{tr } \{\{\int_\omega H^{\frac{1}{2}}(\omega) \ d\omega\}\{\int_\omega H^{\frac{1}{2}}(\omega)d\omega\}S^{-1}\} . \qquad (3.49)$$

For the filter assumption that the noise is white, $S_{vv}(\omega)$=S= constant, the trace of the information matrix is

$$\text{tr } M_f = \int_{\omega_{min}}^{\omega_{max}} \text{tr}\{H \ (\omega) \ S^{-1}\} \ d\omega$$

$$= \int_{\omega_{min}}^{\omega_{max}} \text{tr}\{H^{\frac{1}{2}}(\omega)H^{\frac{1}{2}*}(\omega)S^{-1}\} \ d\omega$$

$$\text{tr } M_f{}^a = \text{tr}\{\{\int_{\omega_{min}}^{\omega_{max}} \{H^{\frac{1}{2}}(\omega)H^{\frac{1}{2}*}(\omega)\} \ d\omega \ S^{-1}\} \quad . \qquad (3.50)$$

Since the product of integrals is larger or equal to the integral of a product,

$$\text{tr } M_f \geq \text{tr } M_f^a \quad . \qquad (3.51)$$

Therefore, the error in the parameter estimate increases if white noise is assumed for an estimation problem with non-white errors. A good measure for the degradation of the maximum likelihood estimator if worst type noise is present and white noise is assumed is,

$$\frac{1}{p} \text{tr } [M_f{}^{-1} M_f{}^{a}] = \text{tr } \{[H^{-\frac{1}{2}} H^{-\frac{1}{2}*}]^{-1} \int_{\omega} H(\omega) \, d\omega\} \frac{1}{p} \quad . \tag{3.52}$$

where p is the number of parameters. The error covariance of the parameter estimate $\hat{\theta}$ increases, therefore, by the noise distribution factor

$$f_{nd} = \frac{p}{\text{tr}\{ [H^{-\frac{1}{2}} H^{-\frac{1}{2}*}]^{-1} \int_{\omega} H(\omega) d\}} \quad . \tag{3.53}$$

3.4.4 Example: First Order System

Let the parameter sensitivity be

$$\frac{\partial h}{\partial \theta}(\omega) = \begin{cases} \dfrac{1}{1+j\omega} & \text{for } 0 < \omega \leq \omega_{max} \\ 0 & \text{for } \omega > \omega_{max} \end{cases} , \tag{3.54}$$

with the noise spectrum

$$S(\omega) = \begin{cases} C & \text{for } \omega < \omega_{w_{max}} \\ \dfrac{C_1}{\sqrt{1+\omega^2}} & \text{for } \omega_{w_{max}} < \omega < \omega_{max} \end{cases} , \tag{3.55}$$

and the covariance

$$\int_0^{\omega_{max}} S(\omega) \, d\omega = R \quad . \tag{3.56}$$

The trace of information matrix is, $\text{tr } M = \int_{\omega} \text{tr } \{\frac{\partial h}{\partial \theta} S_{vv}(\omega)^{-1} \frac{\partial h}{\partial \theta}^{*}\} d\omega$. (3.57)

In the white noise case the spectral distribution is constant, $S(\omega) = \dfrac{R}{\omega_{max}}$.

The trace of the information matrix is, therefore,

$$\text{tr } M_w = \int_{\omega_{min}}^{\omega_{max}} \frac{1}{1+\omega^2} \frac{\omega_{max}}{R} \, d\omega = \frac{\omega_{max}}{R} \quad \text{arctan } \omega_{max} . \tag{3.58}$$

For a mixture of white and non-white noise:

$$S(\omega) = \begin{cases} C & \text{for } 0 \leq \omega \leq \omega_{w_{max}} \\ \dfrac{C\sqrt{1+\omega_{w_{max}}{}^{2}}}{\sqrt{1+\omega^{2}}} & \text{for } \omega_{w_{max}} < \omega \leq \omega_{w_{max}} \end{cases} \qquad (3.59)$$

$$\text{with } C = \frac{R}{\omega_{w_{max}} + \sqrt{1+\omega_{w_{max}}}\ \text{Arsh}[\omega_{max}\sqrt{1+\omega_{w}} - \omega_{w}\ \sqrt{1+\omega_{max}}]} \ .$$

For the white noise section,

$$\frac{1}{C}\int_{0}^{\omega_w} \frac{1}{1+\omega^2}\, d\omega = \frac{1}{C} \arctan \omega_w \ ; \qquad (3.60)$$

non-white section,

$$\frac{1}{C_1}\int_{\omega_w}^{\omega_{max}} \frac{\sqrt{1+\omega^2}}{1+\omega^2}\, d\omega = \frac{1}{C_1}\int_{\omega_w}^{\omega_{max}} \frac{1}{\sqrt{1+\omega^2}}\, d\omega = \frac{1}{C_1}\ \text{Arsh}(\omega)\Big|_{\omega_w}^{\omega_{max}} \ ,$$

$$\text{tr } M_{nw} = \frac{\omega_w + \sqrt{1+\omega_w{}^2}\ \text{Arsh } \omega^*}{R}\ [\arctan \omega_w + \frac{\text{Arsh } \omega^*}{\sqrt{1+\omega_w{}^2}}], \text{ with} \qquad (3.61)$$

$$\omega^* = \omega_{max}\sqrt{1+\omega_w{}^2} - \omega_w\ \sqrt{1+\omega_{max}{}^2} \ .$$

The error covariance of the estimate for the non-white case relative to the white noise case is then $\text{tr}M_{nw}{}^{-1}/\text{tr}M_{w}{}^{-1}$. Figure 3.1 shows this covariance ratio for different degrees of non-whiteness, indicated by the ratio $\omega_{w_{max}}/\omega_{max}$. In this example, the error covariance for the worst type of noise spectrum is 64% larger than the covariance for white noise. For higher order systems non-white noise will degrade parameter estimates even more.

Non-white noise has therefore a significant influence on parameter estimates and should be considered in estimation procedures.

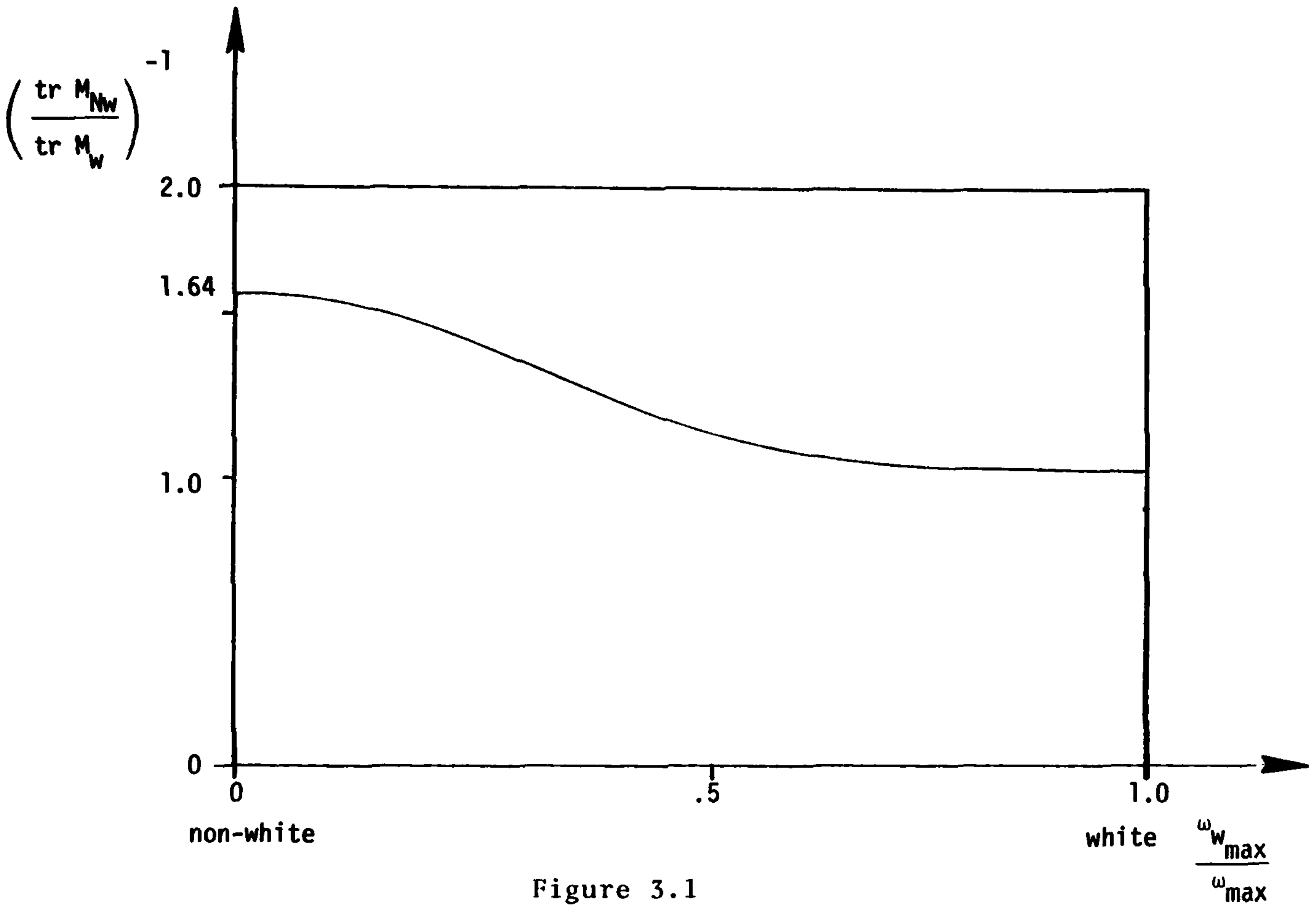


Figure 3.1

DEPENDENCE OF ERROR COVARIANCE OF ESTIMATE ON WHITENESS OF NOISE

(ω_{max} = 10, R = 1)

3.5 ESTIMATION OF CORRELATION

If large data sets are available the distribution function and/or density function may be estimated. In the following, a method for estimating the distribution function is outlined; the density function can be estimated in a similar way.

The estimate $\hat{F}(x)$ of a distribution function $F(x)$ may be described by a parametric form,

$$\hat{F}(x) = f(\alpha,\phi(x)), \tag{3.62}$$

where x is a vector of unknown coefficients and $\Phi(x)$ is a set of independent functions of x, e.g. moment functions.

The coefficients α may then be estimated by minimizing a suitable function of the error $f(\alpha,\phi(x))-F(x)$. In the least square sense, the estimate of $\hat{\alpha}$ of α is,

$$\hat{\alpha} = \arg\min_{\alpha} E[f(x,\phi(x))-F(x)]^2 \tag{3.63}$$

If $f(x, (x))$ is a linear function of α,

$$f(\alpha,\phi(x) = \alpha^T\phi(x) \ , \tag{3.64}$$

the least squares estimate (LSE) of α becomes

$$\hat{\alpha} = [E\{\phi(x)\phi^T(x)\}]^{-1}E\{\phi(x)F(x)\}. \tag{3.65}$$

For numerical evaluation, it suffices generally to replace the estimate by an integral. The estimate of α becomes then,

$$\hat{\alpha} = [\int_R \phi(x)\phi^T(x)dx]^{-1} \int_R \phi(x)F(x)dx, \tag{3.66}$$

where $\Pi dx = dx_i$, and R is the range of x where the estimate is desired. Kashyap/Blaydon (1968) developed a gradient algorithm which sequentially updates estimates α when new data are available.

For the estimation of mixtures of normal or exponential density functions, it is numerically easier to estimate the logarithm of the density function, by either maximizing a regression criteria,

$$\begin{aligned} J_R &\hat{=} E\,[\ln \hat{f}(a,x)] = \int f(x) \ln \hat{f}(a,x)dx, \\ &\int \hat{f}(a,x)\, dx = 1 \end{aligned} \tag{3.67}$$

or, equivalently, minimizing an error or information criteria (Young, Coraluppi, 1970),

$$J_E \hat{=} E[\ln \frac{f(x)}{\hat{f}(a,x)}] = \int f(x) \ln \frac{f(x)}{\hat{f}(a,x)}\, dx \ . \tag{3.68}$$

The estimated density function may now be used to design a maximum likelihood estimator. The sequential correlations of the residuals of this estimator can then be computed and adaptively or iteratively included in the estimator.

3.6 BOUNDING LINEAR ESTIMATOR

In many measurement systems, auto correlations of measurement errors are known, but little information is available about cross correlations between error sources, correlations between system and measurement errors, and sequential correlation (non-white errors). Hence, an estimator that includes a priori knowledge of cross correlations and sequential correlations cannot be designed. Estimators for such systems can be designed with an upper bound on estimation errors.

3.6.1 Bounding Linear Miminum Variance Estimation

Let us assume a system,

$$x_{k+1} = \phi_k x_k + \Gamma_k w_k, \quad k=0,1,\ldots,N-1$$

with output measurements,

$$y_k = H_k x_k + v_k, \quad k=1,2,\ldots,N,$$

where $x_k \varepsilon R^n$; y_k, $v_k \varepsilon R^m$; $w_k \varepsilon R^p$.

Then a bounding estimator provides as estimate $\hat{x}$ of x, such that

$$E[(x_k-\hat{x}_k)(x_k-\hat{x}_k)^T] \leq P_B(k|k).$$

Combining x_o, w, v in a composite state,

$$x^T{}_i := [x_o{}^T, w_o{}^T, v_1{}^T, \ldots, w_{i-1}{}^T, v_i{}^T]$$

and with the composite measurement vector,

$$z_i{}^T := [z_i{}^T, z_z{}^T, \ldots, z_i{}^T],$$

the measurement matrix becomes,

$$H_i: = \left.\underbrace{\begin{bmatrix} H_1\phi_0 & H_1\Gamma_0 & I & 0 & 0 & 0 \\ H_2\phi_0\phi_1 & H_2\phi_1\Gamma_0 & 0 & H_2\Gamma_1 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ H_i\phi_0\cdots\phi_{i-1} & H_i\phi_1\cdots\phi_{i-1}\Gamma_0 & 0 & H_i\phi_2\cdots\phi_{i-1}\Gamma_1 & H_i\Gamma_{i-1} & I \end{bmatrix}}_{n \;+\; (p\,+\,m)i}\right\} mi$$

Then the causal system can be written

$$z_i = H_i \; x_i \; .$$

For w, v random zero mean variables the predicted value for the composite state x_i is

$$E[x_i] = \bar{x}_i = [\bar{x}_0^T, \theta, \ldots, 0]$$

with positive semi-definite covariance

$$R_i = E[(x_i - \bar{x}_i)(x_i - \bar{x}_i)^T]$$

of order n + (p + m)i.

For sequentially and colored noise, the correlation matrix is,

$$R_i = \begin{bmatrix} R_{xx}(0,0) & R_{x_0w_0} & R_{x_0v_1} & R_{x_0w_1} & R_{x_0v_2} \cdots & R_{x_0v_i} \\ R_{w_0x_0} & R_{w_0w_0} & R_{w_0v_1} & R_{w_0w_1} & R_{w_0v_2} \cdots & R_{w_0v_i} \\ R_{v_1x_0} & R_{v_1w_0} & R_{v_1v_1} & R_{v_1w_1} & R_{v_1v_2} \cdots & R_{v_1v_i} \\ \cdot & & & & & \\ \cdot & & \ddots & & & \\ \cdot & & & & & \\ R_{w_{i-1}x_0} & R_{w_{i-1}w_0} & R_{w_{i-1}v_1} & R_{w_{i-1}w_1} & R_{w_{i-1}v_2} \cdots & R_{w_{i-1}v_i} \\ R_{v_ix_0} & R_{v_iw_0} & R_{v_iv_1} & R_{v_iw_1} & R_{v_iv_z} \cdots & R_{v_iv_i} \end{bmatrix}$$

with no sequential correlation,

$$R_{w_j w_k} = R_{w_j v_k} = R_{v_j v_k} = 0 \text{ for } k \neq j \quad ,$$

the covariance matrix R_i becomes block diagonal and the estimator can be written in a Kalman filter type sequential updating form. With $R_{vw} = 0$, no correlation between the noise sources, R becomes diagonal. The estimator becomes now a regular Kalman filter for Gaussian white noise. If R_i is known, the linear minimum variance estimate $\hat{x}_i$ of x_i is,

$$\hat{x}_i = \bar{x}_i + R_i H_i^T [H_i R_i H_i^T]^{-1} [z_i - H_i \bar{x}_i] \quad ,$$

with error covariance,

$$E[(x_i - \hat{x}_i)(x_i - \hat{x}_i)^T] = R_i - R_i H_i^T [H_i R_i H_i^T]^{-1} H_i R_i \quad .$$

For an estimator of the form,

$$\hat{x}_i = \bar{x}_i + K_i (z_i - H_i \bar{x}_i) \quad ,$$

the error covariance of the estimate is

$$E[(x_i - \hat{x}_i)(x_i - \hat{x}_i)^T] = [I - K_i H_i] R_i [I - K_i H_i]^T \quad .$$

For an upper estimate $Q_i \geq R$

$$[I - K_i H_i] R_i [I - K_i H_i]^T \leq [I - K_i H_i] Q_i [I - K_i H_i]^T$$

(pos. semidefinite if $m \geq n$). With the minimum variance gain

$$K_i = Q_i H_i^T [H_i Q_i H_i^T]^{-1} \quad ,$$

the bounding estimator is

$$\hat{x}_i = \bar{x}_i + Q_iH_i^T[H_iQ_iH_i^T]^{-1}[z_i - H_i\bar{x}_i]$$

$$E[(x_i - \hat{x}_i)(x_i - \hat{x}_i)^T] \leq Q_i - Q_iH_i^T[H_iQ_iH_i^T]^{-1}H_iQ_i \quad .$$

Since this estimate is a bounding estimator for every $Q \geq R$, Q can be chosen such that

$$Q = \text{diag } Q ,$$

and a sequentially updating bounding estimator can be designed. The residuals of the bounding estimator may be tested for non-white and correlated noise. This information can then be used to tighten the bounds of the estimator.

3.6.2 Bounding Likelihood Estimators

In a similar way, for the bounding linear minimum variance estimator bounding likelihood estimates can be determined by deweighting the innovation sequence of a sequential estimator. The negative log-likelihood function becomes then

$$\text{NLLF} = (y - H\hat{x})^T Q^{-1}(y - H\hat{x}) + \ln|R| ,$$

where $Q = \text{diag } Q \geq R$ is an upper bound of the error covariance. This change can be incorporated in existing state and parameter estimation algorithms in order to get a bounding estimator for correlated and non-white noise.

3.7 CONCLUSIONS

The estimation problem for correlated as well as non-white noise was analyzed. Estimation for a priori known correlations are shown. The degradation of a filter due to non-whiteness of the noise was shown for a linear filter that assumes white noise. Estimators that give unbiased estimates for correlated and non-white noise, with bounds on the estimate error covariance, are shown. Further research is necessary to deal with the problem of non-white noise due to non-linearities and unmodeled states.

IV. CONCLUSIONS AND RECOMMENDATIONS

4.1 CONCLUSIONS

Analysis of test results indicates that the measurement and process noise is significantly non-white and non-Gaussian. Some analyses indicate that 10% to 15% of the data points may deviate significantly from non-Gaussian distribution. In addition, numerous sources lead to non-white noise.

These errors effect both the accuracy of state and parameter estimates as well as the estimation of accuracy levels. In this report, techniques have been developed to treat systems with non-white and non-Gaussian noise. These techniques provide good estimates under given whiteness and Gaussianess conditions. The procedures are simple and can be easily incorporated in the standard maximum likelihood and model structure determination methods.

4.2 RECOMMENDATIONS

Algorithms used in system identification and parameter estimation should be modified to include the effects of non-white and non-Gaussian noise. Such modifications are necessary to significantly improve the accuracy with which parameters/states are estimated.

REFERENCES

1. G.B. Airy, Letter from Professor Airy, Astronomer Royal, to the editor, Astronomical Journal, 4, pp. 137-138, 1856.

2. A. Birnbaum, E. Laska, "Optimal Robustness: A general Method with Applications to Linear Estimates of Location," J. American Statistical Assoc., 62, pp. 1230-1240, 1967.

3. A.E. Bryson, L.J. Henrikson, "Estimation Using Sampled Data Containing Sequentially Correlated Noise," Journal Spacecraft and Rockets, 5, 1968, pp. 662-666.

4. A.E. Bryson, Y.C. Ho, "Applied Optimal Control," Ginn and Company, 1969.

5. A.E. Bryson, D.E. Johansen, "Linear Filtering for Time-Varying Systems Using Measurements Containing Colored Noise," IEEE Trans. Automatic Control, AL-10, 1965, p. 4.

6. D.R. Cos, P.A.W. Lewis, The Statistical Analysis of Series of Events, Methuen, London 1966.

7. P.S. de Laplace, Deuxieme Supplement a la Théorie Analytique des Probabilités, Courcier, Paris, 1818.

8. F.Y. Edgeworth, "Problems in Probabilities," Philosophical Magazine, 22, pp. 371-384, 1886.

9. R.L. Ellis, "On the Method of Least Squares," Transactions of the Cambridge Philosophical Society, pp. 204-219, 1844.

10. G.T. Fechner, "Ueber den Ausgangswert der kleinsten Abweichungssumme, dessen Bestimmung, Verwendung und Verallgemeinerung," Abhandlungen der Mathematisch-Physikalischen Classe der Königlich Sächsischen Gesellschaft der Wissenschaften, Leipzig, pp. 1-76, 1878.

11. J. R. Fisher, E.B. Stear, "Optimal Nonlinear Filtering for Independent Increment Processes," Parts I and II, IEEE Transactions on Information Theory, IT-3, pp. 558-578, 1967.

12. H.W. Freedman, "The Little Variable Factor. A Statistical Discussion of the Reading of Seismograms," Bulletin of the Seismology Society of America, 56, pp. 593-604, 1966.

13. J.L. Gastwirth, H. Rubin, "The Behavior of Robust Estimators on Dependent Data," Mimeograph Series No. 197, Purdue University, Dept. of Statistics, 1969.

REFERENCES (Continued)

14. W.E. Hall, Jr., N.K. Gupta, R.G. Smith, "Identification of Aircraft Stability and Control Coefficients for the High Angle-of-Attack Regime," Systems Control, Inc., Technical Report No. 2, 1974.

15. F.R. Hampel, "A General Qualitative Definition of Robustness," Annals of Math. Statistics, 42 pp. 1887-1896, 1971.

16. F.R. Hampel, "Robust Estimation: A Condensed Partial Survey," Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, pp. 87-104, 1973.

17. L.J. Henrikson, "Sequentially Correlated Measurement Noise with Application to Inertial Navigation," J. of Spacecraft and Rockets, 5, 1968.

18. A. Høyland, "Robustness of Wilcoxon Estimates of Location Against a Certain Dependence," Annals of Math. Statistics, 39, pp. 1191-1201, 1968.

19. P.J. Huber, "Robust Estimation of a Location Parameter," Annals of Math. Statistics, pp. 73-101, 1964.

20. P.J. Huber, "Robust Statistics: A Review," Annals of Math. Statistics, 43, pp. 1041-1067, 1972.

21. P.J. Huber, Robust Statistical Procedures, Society for Industrial and Applied Mathematics, Philadelphia, 1977.

22. B.N. Jain, "Bounding Estimators for Systems with Colored Noise," IEEE Transactions on Automatic Control, pp. 365-368, 1975.

23. A.H. Jazwinski, "Stochastic Processes and Filtering Theory," Academic Press, 1970.

24. R.E. Kalman, "New Methods in Wiener Filtering Theory," Proceedings of Symp. Eng. Appl. Random Function Theory and Probability, (J.L. Bogdanoff, F. Koziu, eds.) Wiley, 1963.

25. R.L. Kashyap, C.C. Blaydon, "Estimation of Probability Density and Distribution Functions, "IEEE Transactions on Information Theory, IT-14, pp. 549-556, 1968.

26. A.M. Legendre, "On the Method of Least Squares," (1805), in A Source Book in Mathematics, Dover, New York, 1959.

REFERENCES (Continued)

27. J.I. McCool, "Inference on Weibull Percentiles and Shape Parameter from Maximum Likelihood Estimates," IEEE Transactions on Realiability, R-19, pp. 2-9, 1970.

28. J.A. McFadden, "The Entropy of a Point Process," SIAM J. of Applied Mathematics, 13, pp. 988-994, 1965.

29. R. D. Martin, S.C. Schwartz, "On Mixture, Quasi-Mixture, and Nearly Normal Random Processes," Annals of Mathematical Statistics, 43, pp. 948-967, 1972.

30. R.K. Mehra, "On the Identification of Variances and Adaptive Kalman Filtering," IEEE Trans. on Autom. Control, AC-15, pp. 175-184, 1970.

31. R.K. Mehra, A.E. Bryson, "Linear Smoothing Using Measurements Containing Correlated Noise with an Application to Inertial Navigation," IEEE Trans. on Automatic Control, AC-10, 1968.

32. H.M. Merrill, "Bad Data Suppression in State Estimation, with Application to Problems in Power," Thesis, MIT, Cambridge, Mass., June 1972.

33. S. Newcomb. "A Generalized Theory of the Combination of Observations so as to obtain the Best Results," American Journal of Mathematics, pp. 343-366, 1886.

34. G.E. Noether, "Wilcoxon Confidence Intervals for Location Parameters in the Discrete Case," J. American Statistical Association, 62, pp. 184-188, 1967.

35. P. Papantoni-Kazakos, R.M. Gray , "Robustness of Estimators on Stationary Observation," to be published in J. of American Mathematical Society, 1979.

36. R.W. Parks, "Efficient Estimation of a System of Regression Equations when Disturbances are Both Serially and Contemporaneously Correlated," J. American Statistical Association, pp. 500-509, 1967.

37. B. Peirce, "Criterion for the Rejection of Doubtful Observations," Astronomical J., 2, pp. 161-163, 1852.

38. Y.V. Prokhorov, "Convergence of Random Processes and Limit Theorems in Probability Theory," Theoretical Probability Applications, 1, pp. 157-214, 1956.

REFERENCES (Continued)

39. I. Rubin, "Regular Point Processes and their Detection," IEEE Transactions on Information Theory, IT-18, pp. 547-557, 1972.

40. I. Rubin, "Reduced-Memory Likelihood Processing of Point Processes," IEEE Transactions on Information Theory, IT-20, pp. 729-738, 1974.

41. D.L. Snyder, "Filtering and Detection for Doubly Stochastic Poisson Processes," IEEE Transactions on Information Theory, IT-18, pp. 91-102, 1972.

42. S.M. Stigler, "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920," Journal of the American Statistical Association, pp. 872-879, 1973.

43. J.W. Tukey, "A Survey of Sampling from Contaminated Distributions," in Contributions to Probability and Statistics, Stanford University Press, pp. 448-485, 1960.

44. G.L. Wise, A.P. Traganities, J.B. Thomas, "The Estimatiom of a Probability Density Function from Measurements Corrupted by Poisson Noise," IEEE Transactions on Information Theory, IT-23, pp. 764-766, 1977.

45. T.Y. Young, G. Coraluppi, "Stochastic Estimation of a Mixture of Normal Density Functions Using an Information Criterion," IEEE Transactions on Information Theory, IT-16, pp. 258-263, 1970.

APPENDIX A

MAXIMUM LIKELIHOOD ESTIMATION IN THE PRESENCE OF POISSON NOISE

INTRODUCTION

In many problems of interest, the state equations are governed by

$$\dot{x} = Fx + Gu \qquad t_o \leq t \leq T \tag{1}$$

and the measurements are arrival times based on Poisson processes. The probability of arrival in an interval t to $t+\Delta t$ depends on a linear combination of the state variables

$$p(t, t+\Delta t) = Hx(t)\Delta t \tag{2}$$

Consider a scalar arrival sequence described by the above process. Given the arrival times $t_1 < t_2 < t_3 \ldots t_N$, the problem is to estimate unknown parameters in F, G, and H.

This solution will be useful in low-intensity image processing (where photon release follows Poisson process) and medical imaging with radioactive tracers.

MAXIMUM LIKELIHOOD ESTIMATION

The probability density functin for arrival times t_i, $i=1,2, \ldots N$ follows the equation

$$\mathscr{L}(\theta | t_i) = f(t_1,t_2 \ldots t_N) = \prod_{i=1}^{N} \mu_i \, e^{-\mu_i} \tag{3}$$

$$\mu_i = \int_{t_{i-1}}^{t_i} Hx(t)dt \tag{4}$$

The negative log likelihood function (NLLF) of parameters given the arrival times is

$$J = - \ln[\mathscr{L}(\theta | t_i)] = \sum_{i=1}^{N} [\mu_i - \ln \mu_i] \tag{5}$$

The first gradient of the NLLF is

$$\frac{\partial J}{\partial \theta} = \sum_{i=1}^{N} [1-\frac{1}{\mu_i}] \frac{\partial \mu_i}{\partial \theta} \tag{6}$$

$$\frac{\partial \mu_i}{\partial \theta} = \int_{t_{i-1}}^{t_i} [\frac{\partial \mu}{\partial \theta} x(t) + H \frac{\partial x(t)}{\partial \theta}] \, dt \tag{7}$$

The second gradient is obtained as follows

$$\frac{\partial^2 J}{\partial \theta^2} = \sum_{i=1}^{N} [\frac{1}{\mu_i^2} \frac{\partial \mu_i}{\partial \theta} \left(\frac{\partial \mu_i}{\partial \theta}\right)^T + (1-\frac{1}{\mu_i}) \frac{\partial^2 \mu_i}{\partial \theta^2}] \tag{8}$$

The second gradient may be approximated as follows

$$\frac{\partial^2 J}{\partial \theta^2} \simeq \sum_{i=1}^{N} \frac{1}{\mu_i^2} \left(\frac{\partial \mu_i}{\partial \theta}\right)\left(\frac{\partial \mu_i}{\partial \theta}\right)^T \tag{9}$$

For a single count at t_1, one parameter may be estimated by setting

$$\mu_1 = \int_{t_o}^{t_1} Hx(t)dt = 1 \tag{10}$$

The second gradient matrix has a complex distribution. Its expected value may be approximated by

$$E(\frac{\partial^2 J}{\partial \theta^2}) \approx \sum_{i=1}^{N} \left(\frac{\partial \bar{\mu}_i}{\partial \theta}\right)\left(\frac{\partial \bar{\mu}_i}{\partial \theta}\right)^T \tag{11}$$

where t_i's are such that

$$\bar{\mu}_i = \int_{t_{i-1}}^{t_i} Hx(t)dt = 1 \tag{12}$$

COUNTS OVER A SAMPLE PERIOD

The arrival times are difficult to use in estimation when there are many occurrences over the time period of observations. The estimation may be based on the number of counts over a set of sample periods. Let $y_1, y_2 \ldots y_N$ be the number of counts over periot t_o, $t_o+ t, \ldots t_o+N\ t$. The likelihood function of parameters θ based on measurements y_i, $i=1,2\ldots N$ is

$$(\theta|y) = f(y|\theta) = \prod_{i=1}^{N} (\mu_i^{y_i} e^{-\mu_i})/(y_i!) \tag{13}$$

$$\mu_i = \int_{t_o+(i-1)\Delta}^{t_o+i\Delta} Hx(t)dt \tag{14}$$

The negative log-likelihood function is

$$J = -\ln[\mathscr{L}(\theta|y)] = \sum_{i=1}^{N} [-y_i \ln(\mu_i) + \mu_i + \ln(y_i!) \tag{15}$$

The first and the second gradients are

$$\frac{\partial J}{\partial \theta} = \sum_{i=1}^{N} \left(1 - \frac{y_i}{\mu_i}\right) \frac{\partial \mu_i}{\partial \theta} \tag{16}$$

$$\frac{\partial^2 J}{\partial \theta^2} = \sum_{i=1}^{N} \left[\frac{y_i}{\mu_i^2} \left(\frac{\partial \mu_i}{\partial \theta}\right)\left(\frac{\partial \mu_i}{\partial \theta}\right)^T + \left(1 - \frac{y_i}{\mu_i}\right) \frac{\partial^2 u_i}{\partial \theta^2}\right] \tag{17}$$

An approximation to the second gradient is

$$\frac{\partial 2_J}{\partial \theta^2} \approx \sum_{i=1}^{N} \frac{y_i}{\mu_i^2} \left(\frac{\partial \mu_i}{\partial \theta}\right)\left(\frac{\partial \mu_i}{\partial \theta}\right)^T \tag{18}$$

A Newton-Raphson step may be taken in an iterative procedure as follows

$$\theta_{k+1} = \theta_k - \left(\frac{\partial^2 J}{\partial \theta^2}\right)^{-1} \left(\frac{\partial J}{\partial \theta}\right) \tag{19}$$

This procedure may be continued until convergence occurs.

The expected value of the second gradient matrix is

$$E \approx \sum_{i=1}^{N} \frac{1}{\mu_i} \left(\frac{\partial \mu_i}{\partial \theta}\right) \left(\frac{\partial \mu_i}{\partial \theta}\right)^T \tag{20}$$

There is no approximation in the above equation.