Analytical Redundancy Approach for Fault Detection in Electro-Hydraulic Servo-Positioning Systems Using A Nonlinear Observer and Sequential Test of Wald

by

Hameedullah Khan

A Thesis Submitted to the Faculty of Graduate Studies of the

University of Manitoba in Partial Fulfillment of the Requirements

of the Degree of

Master of Science

Department of Mechanical and Industrial Engineering

University of Manitoba

Winnipeg, Manitoba

CANADA R3T 2N2

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

ANALYTICAL REDUNDANCY APPROACH FOR FAULT DETECTION IN ELECTRO-HYDRAULIC SERVO-POSITIONING SYSTEMS USING A NONLINEAR OBSERVER AND SEQUENTIAL TEST OF WALD

BY

HAMEEDULLAH KHAN

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

of

MASTER OF SCIENCE

Abstract

This thesis presents a scheme for condition monitoring and fault detection in hydraulic servo positioning systems using analytical redundancy. The work contributes to fault detection and isolation (FDI) of nonlinear systems. In this context, a fault is defined as any kind of malfunction in the dynamic system that leads to an unacceptable anomaly in overall system performance. Failures inevitably occur at the most inconvenient times creating technical, organizational and financial restrictions that could be minimized with the application of appropriate condition monitoring and fault detection techniques. For example, repair costs due to fluid losses in heavy equipment alone can result in hundreds of thousands of dollars in losses apart from those due to the failure of other components.

The approach taken in this thesis is to reconstruct the output of a system by an observer. The comparison between the estimated and the measured values of the output is permitted by using an average sequential test of Wald to detect the failure. Sequential testing takes into account the past evolution of the tested variable. One of the following three decisions can be chosen at any stage of the experiment: (1) to accept the hypothesis, (2) to reject the hypothesis, and (3) to continue the experiment by making additional observations. Wald's sequential test is a method of statistical inference whose characteristic feature is that the number of observations required by the procedure is not determined in advance of the experiment. This test frequently results in a savings of about 50% in the number of observations over the most efficient test procedure based on a fixed number of observations.

Velocity output from the system is taken as input to the observer. Residuals are taken from the difference of the actual and the estimated velocities. The cumulative sum of the residuals is then used in a sequential test of Wald to determine the occurrence of fault. Simulation studies are conducted for cross-port leakage in the actuator, external leakage from the actuator, faults due to compliance variation, faults due to incorrect supply

pressure, and sensor faults. Experiments are conducted for faults due to incorrect supply pressures and sensor faults. Both simulation and experimental results show the promise and potential of the proposed method.

Acknowledgments

First of all, I would like to express my special gratitude to my advisor, Dr. Nariman Sepehri, who was a great source of inspiration for me, with his hard-working, caring personality and academic knowledge. There were many stressful occasions in which he helped me enormously.

I am thankful to my great friend Dr. Seraphin Chally Abou for his help, support and encouragement in my research. I would also like to thank the members of my advisory committee, Dr. Onyshko and Dr. Balakrishnan, for the careful reading of my thesis and useful comments.

I would also like to thank all my friends, particularly Dr. Amir Ali Akbar Khayyat and Navid Niksefat for their encouragement.

Finally, I reserve my warmest gratitude for my beloved and respected parents for their sacrifices, encouragement, and for always being supportive of my decisions.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

Hydraulic systems are used in most industries where heavy objects are manipulated or large forces are exerted on their surroundings. Examples include pick and place robots for industrial applications, positioning of aircraft control surfaces, flight simulators, road simulators, excavators and feller bunchers. Hydraulic systems consist of components such as servovalves, actuators and pumps whose dynamic characteristics are complex, nonlinear and time-varying. For example, as the operating temperature changes, the temperature sensitive parameters such as density, viscosity and bulk modulus of the working fluid change accordingly. For hydraulic systems that operate for long periods of time or in high temperature environments, the effect of these changes cannot be ignored. Hydraulic systems are also subjected to many faults such as actuator faults, leakages, and pump pressure faults. A fault is defined as an undesirable deviation of a characteristic property, which leads to the inability to fulfil an intended purpose (Iserman, 1984). A good fault detection and condition monitoring scheme can save a lot of time and cost. There are many parameters that can be monitored in typical fluid power systems. Measurements may be either of the detailed micro-type such as component wear and spool displacement, or measurement may be of the macro-type resulting from micro-type problems. The type of instrumentation used and the cost depend upon the application and are related to the value of keeping the particular component or system in operation. The more advanced techniques require complex signal processing. Visual inspection of both components and system is perhaps the most basic approach and often initiates automatic condition monitoring. Regular inspection for fluid leaks and audible changes in noise level can often avoid a serious failure (Watton et al., 1987).

It is of great importance to have an efficient operation of hydraulic systems due to their large variety of implementation. It is also necessary to have a condition monitoring and fault detection scheme to detect the fault as it occurs. There are several types of faults which can occur in a hydraulically actuated system. A drop in pump pressure could cause the stalling of the actuator against the load. In an aerospace application where flight control surfaces are moved by a hydraulic actuator, stalling of the actuator could have disastrous consequences. Lack of pump pressure can result in extreme cases, from increased leakage in the pump due to wear or a burst in a supply line. A relief valve failure can also cause the pump pressure to drop. Similarly, a relief valve that remains closed can cause the pump pressure to rise substantially and lead to destruction and failure of the system due to overpressure.

Another common source of a fault is the leakage from actuators and supply lines. A hydraulic actuator can suffer from two different types of leakages. The first is leakage at the piston seal or internal flow loss. When leakage occurs across a piston seal, no fluid is lost from the system, but the motion of the actuator is affected since only a portion of the flow passing through the valve is available to move the actuator. The second type of leakage is one at the shaft seal or external flow loss. Fluid is lost from the system when external leakage occurs. If external leakage is not detected, the loss of fluid from the system could result in a substantial drop in line pressures and the hydraulic system will eventually cease to operate. Watton et al. (1987) have investigated expert system and neural network approaches to diagnosing internal and external leakages in hydraulic actuators and lines.

The third common cause of malfunction and failure in a hydraulic system is particle, air and water contamination of the hydraulic fluid. Since the bulk modulus of air is much smaller than that of hydraulic fluid, any air trapped in the system will result in a significant reduction in the effective bulk modulus of the fluid. Similarly, the presence of water in the system will manifest itself as an increase in the effective bulk modulus of the fluid. The natural frequency of the entire system is a function of the effective bulk modulus. In a closed-loop control system, a change in the natural frequency of the system can have serious implications on the overall system performance. Thus, changes in the

effective bulk modulus should be monitored in high performance closed-loop control systems.

Fault Detection and Isolation (FDI) is the basic technology to detect the occurrence of a fault and determine its cause. Reconfiguration of controls is used to maintain the performance. This can involve sensor signals, the use of actuators and/or revised control algorithms.

In this study we introduce a methodology of fault detection based on an analytical redundancy approach by employing a nonlinear observer and a statistical sequential test. The design of a system that produces an approximation to the state vector in a deterministic setting is called an observer (Luenberger, 1971). The general procedure for fault detection first generates the so-called residuals (i.e., fault-accentuated functions) before proceeding to detection and reconfiguration (Adjallah et al, 1994). Although various FDI methods have been developed, most of them are for linear systems. Since most practical dynamic systems display nonlinearity, the linear model based fault detection methods can only work well in a small region around an operating point (Yu, 2000). Therefore there is a great need for developing FDI methods for nonlinear systems, particularly for fluid power systems as they are widely used in many industrial applications. In order to achieve fault detection and localization for a wide class of nonlinear systems subjected to bounded nonlinearity, a nonlinear observer is usually used (Kudva, 1980). Yu (2000) proposed a robust minimal order state observer for bilinear systems with unknown inputs and applied it to a hydraulic system. Crowther et al. (1998) presented a neural network approach for the fault diagnosis of a hydraulic actuator circuit.

1.2 Objectives and Scope of This Thesis

In this thesis, a fault detection methodology for an electro-hydraulic servo-positioning system using an analytical redundancy approach is presented. Different stages in achieving this goal are as follows: (i) a nonlinear observer is employed for residual generation, (ii) the input signal and the velocity as output of the system are used as inputs to the observer, (iii) residuals, i.e., the differences between the estimated and the

measured velocities, are generated for fault detection strategy, and (iv) the residuals are evaluated via sequential test of Wald, a reliable method to detect the occurrence of a fault.

The organization of this thesis is as follows. Basic concepts of fault detection and isolation are discussed in Chapter 2. Since this thesis presents fault detection using analytical redundancy (model based fault detection), classification of analytical and physical redundancy are also discussed. In Chapter 3, detailed concepts of analytical redundancy are discussed. In any fault detection methodology, it is desired that the algorithm be sensitive to the faults, yet be insensitive to disturbances, noise and modelling errors. Thus various kinds of disturbances and noises are analyzed in Chapter 3. Chapter 4 discusses the fundamental observer based residual generation as this is the primary method adopted in this work for residual generation. Linear and nonlinear observers are discussed in detail. Residual generation is important for fault detection as it provides the knowledge of the state of the system. However, it does not provide sufficient information about the detection of the fault unless it is evaluated by a statistical test for decision-making. There are many statistical methods for fault detection which can be used according to their availability and the nature of the system. These methods are discussed in Chapter 5. Sequential test of Wald, which is used in this work for analyzing the residuals to make a decision for fault detection, is discussed in detail. Chapters 6 and 7 consist of the mathematical modelling of the system and design and analysis of the nonlinear observer. Chapter 7 also includes fault detection methodology adopted in this work. Simulation and experimental studies are presented in Chapter 8 for various faults such as incorrect pump pressure faults, internal and external leakages, changes in the bulk modulus and sensor faults. Experimental results are only presented for incorrect pump pressure and sensor faults. Conclusions are presented in Chapter 9.

2 Fault Detection and Isolation

2.1 Classification of Faults

A fault is defined as an undesirable deviation of a characteristic property which leads to the inability to fulfil an intended purpose. Faults are generally described by deterministic time-functions which are not known. Important special cases are the jump-fault (step function) and the drift-fault (ramp function). It is assumed that faults are not present initially in the system but arrive at some later time, with both their magnitude and arrival time being unknown. It is worthwhile to mention that any noise originating from the plant or from sensors and actuators, is considered random with zero mean (any nonzero mean is handled as a fault or disturbance). Faults are usually classified into the following categories:

(i) Additive faults: these are unknown inputs acting on the plant, which are normally zero and which, when present, cause a change in the plant outputs independent of the known inputs. Such faults best describe plant leaks and loads. The distinction between additive faults and disturbances is subjective: faults are those unknown inputs that we wish to detect and isolate while disturbances are nuisances that we wish to ignore.

(ii) Multiplicative faults: these are changes (abrupt or gradual) in some plant parameters. They cause changes in the plant outputs that depend also on the magnitude of the known inputs. Such faults best describe the deterioration of the plant equipment, such as surface contamination, clogging or partial or total loss of power. Modelling errors are discrepancies between the model (model parameters) and the true system. They may have been present since the origin of the system, or may arise due to operating-point changes. Modelling errors are nuisances, the effect of which we want to suppress. They may be considered as multiplicative disturbances, in contrast to multiplicative faults which are also discrepancies between the model and the true system, but which we wish to detect.

(iii) Sensor faults: these are discrepancies between the measured and the actual values of the individual plant variables. These faults are usually considered additive (independent of the measured magnitude), though some sensor faults such as sticking or complete failure may be better characterized as multiplicative.

(iv) Actuator faults: these are discrepancies between the input command of an actuator and its actual output. Actuator faults are usually handled as additive although some kinds of faults such as complete failure, may be better described as multiplicative.

2.2 Fault Detection and Diagnosis

The indication that something is wrong in a monitored system is called fault detection, whereas the determination of the exact location of the faulty component is called fault isolation. The determination of the magnitude of the fault is referred to as fault identification (Gertler, 1998). The isolation and identification tasks together are referred to as fault diagnosis. While detection is an absolute must in any practical system and isolation is almost equally important, fault identification may not justify the extra effort it requires. Therefore, most practical systems contain only the fault detection and isolation stages and are referred to as FDI systems. As well, in many texts, diagnosis is used simply as a synonym to isolation.

Usually the fault detection and diagnosis tasks take place on-line. These two tasks may be performed in parallel or sequentially. In some diagnostic systems, a single decision conveys not only the fact that a fault is present but also its location. In other systems, the detection task is running continuously while the diagnostic task is triggered only upon the detection of the presence of a fault.

The faults we are dealing with may arise in the basic technological equipment or in its measurement and control instruments, probably sensors and actuators. They may represent performance deterioration, partial malfunctions or total breakdowns. From the point of view of diagnosis, it is of interest to know how a particular fault affects the plant outputs, whether in an additive manner or multiplicative manner.

In most practical situations, fault diagnosis needs to be performed in the presence of disturbances, noise and modelling errors. These interfere with the diagnosis of faults and may lead to false alarms and/or misclassification. Therefore, the diagnostic algorithm

needs to be designed such that it: (i) is made insensitive to the disturbances, (ii) includes mechanisms to suppress the effects of noise, (iii) is robust with respect to modelling errors, and (iv) maintains sufficient sensitivity with respect to faults.

2.2.1 Detection Performance

The detection performance of the diagnostic technique is characterised by a number of important and quantifiable benchmarks:

Fault sensitivity: the ability to detect faults of reasonably small size.

Reaction speed: the ability to detect faults with reasonably small delay after their occurrence.

Robustness: the ability to operate in the presence of noise, disturbances and modelling errors, with few false alarms.

Fault sensitivity, reaction speed and robustness arise from an interplay between faults on one hand and, noise, disturbances, and modelling errors on the other. They are affected by the design of the detection algorithm. In most cases, there are design trade-offs between the various properties.

2.2.2 Isolation Performance

The ability of the diagnostic system to distinguish faults depends upon the physical properties of the plant, the size of faults, noise level, disturbances and modelling errors, and on the design of the algorithm. Multiple simultaneous faults are, in general, more difficult to isolate than single faults. Also, the interplay between faults, disturbances, noise and modelling errors may lead to uncertain or incorrect isolation decisions. Furthermore, some faults may be non-isolable from one another because they act on the physical plant in an undistinguishable way.

2.2.3 Classification

The methods of fault detection and diagnosis may be classified into two major groups- those which do not utilize the mathematical model of the plant (model-free) and those which do (model-based).

- **Model-Free Methods**

Model-free methods range from physical redundancy and special sensors to limit-checking and spectrum analysis to logical reasoning.

Physical Redundancy

In this approach, multiple sensors are installed to measure the same physical quantity. Any serious discrepancy between the measurements indicates a sensor fault. With only two parallel sensors, fault isolation is not possible. With three sensors, a voting scheme can be formed which isolates the faulty sensor. Physical redundancy involves extra hardware cost and extra weight, the latter representing a serious concern in aerospace applications.

Special Sensors

Special sensors may be installed explicitly for detection and diagnosis. These may be limit sensors in hardware. Other special sensors may measure some fault-indicating physical quantity such as sound, vibration and elongation.

Limit Checking

In this approach, which is widely used in practice, plant measurements are compared by computer to preset limits. Exceeding the threshold indicates a fault situation. In many systems, there are two levels of limits, the first serving for pre-warning while the second

triggers an emergency reaction. Limit checking may be extended to monitoring the time-trend of selected variables. While simple and straightforward, limit checking approach suffers from two serious drawbacks:

1. Since the plant variables may vary widely due to normal input variations, the test thresholds need to be set quite conservatively.
2. The effect of a single component fault may propagate to many plant variables, setting off a confusing multitude of alarms and making isolation extremely difficult.

Spectrum Analysis

Spectrum analysis of plant measurements may also be used for detection and isolation. Most plant variables exhibit a typical frequency spectrum under normal operating conditions. Any deviation from this is an indication of abnormality. Certain types of faults may even have their characteristic signature in the spectrum, facilitating fault isolation.

Logic Reasoning

Logic reasoning forms a broad class which is complementary to the methods outlined above, in that it is aimed at evaluating the symptoms obtained by the detection hardware or software. The simplest technique consists of trees of logical rules of the "IF-symptom-AND-symptom-THEN-conclusion" type. Each conclusion can, in turn serve as symptom in the next rule, until the final conclusion is reached. The system may process the information presented by the detection hardware/software or may interact with a human operator, inquiring from him/her about particular symptoms and guiding him/her through the entire logical process.

- **Model-Based Methods**

Model-based fault detection and diagnosis methods utilize explicit mathematical models of the monitored plants. A system's natural mathematical description is in the form of

differential equations, or equivalent transformed representation. However, the monitoring computers operate in a sampled fashion, using sampled data. Therefore, it is customary and practical to describe the monitored plant in discrete time, by a set of difference equations, or equivalents.

Most of the model-based fault detection and diagnosis methods rely on the concept of analytical redundancy. In contrast to physical redundancy, whereby measurements from parallel sensors are compared to each other, sensory measurements are compared to analytically computed values of the respective variable. Such computations use present and/or previous measurements of other variables and the mathematical plant model describing their nominal relationship to the measured variable. The resulting differences, called residuals, are indicative of the presence of faults in the system. Another class of model-based methods relies directly on parameter estimation.

The generation of residuals needs to be followed by a residual evaluation scheme, in order to arrive at detection and isolation decisions. Because of the presence of noise and modelling errors, the residuals are never zero, even if there is no fault. Therefore, the detection decision requires testing the residuals against thresholds. The thresholds are determined empirically or by theoretical considerations.

To facilitate fault isolation, the residual generators are usually designed for isolation enhanced residuals, exhibiting structural or directional properties. The isolation decisions can then be obtained in a structural (Boolean) or directional (geometric) framework, with or without the inclusion of statistical elements.

The residuals, generated to indicate faults, may also react to the presence of noise, disturbances and modelling errors. Desensitizing the residuals to these sources is an important aspect in the design of detection and diagnosis algorithms. In particular:

(i) In order to deal with the effects of noise, the residuals may be filtered and statistical techniques may be applied to their evaluation. The latter may be hampered by insufficient information concerning the statistical properties of the noise and the noise-transfer dynamics of the plant.

(ii) Disturbance decoupling may be built into the design of the residual generator, but it competes with isolation enhancement for the available design freedom.

(iii) Robustness in the face of modelling errors is the most fundamental problem in model-based fault detection and isolation. Several methods are available which usually rely on some sort of optimization. Unfortunately, this problem does not lend itself to easy solutions and the known techniques are effective only under limited circumstances. There are four somewhat overlapping approaches to residual generation in a model-based fault detection and isolation technique:

a. Kalman Filter

The error prediction of the Kalman filter can be used as a fault detection residual. Its mean is zero if there is no fault or disturbance and becomes nonzero in the presence of faults. Since the innovation sequence is white, statistical tests are relatively easy to construct (Gertler, 1998). However, fault isolation is somewhat awkward with the Kalman filter. One needs to run a bank of matched filters, one for each suspected fault and for each possible arrival time, and check which filter output can be matched with the actual observations.

b. Diagnostic Observers

Observer innovations also qualify as fault detection residuals. "Unknown input" design techniques may be used to decouple the residuals from a limited number of disturbances. The residual sequence is coloured, which makes statistical testing somewhat complicated. The freedom in the design of the observer can be utilized to enhance the residuals for isolation. The dynamics of the fault detection can be controlled, within certain limits, by placing the poles of the observer.

c. Parity (consistency) Relations

Parity relations are rearranged direct input-output model equations, subjected to a linear dynamic transformation. The transformed residual serves for detection and isolation. The residual sequence is coloured, as in the case of observers. The design freedom provided

by the transformation can be used for disturbance decoupling and fault isolation enhancement. Also, the dynamics of the response can be assigned within the limits imposed by the requirements of causality stability.

d. Parameter Estimation

Parameter estimation is a natural approach to the detection and isolation of parametric (multiplicative) faults. A reference model is obtained by first identifying the plant in a fault-free situation. Next, the parameters are repeatedly re-identified on-line. Deviations from the reference model serve as a basis for detection and isolation. Parameter estimation may be more reliable than the analytical redundancy methods, but it is also more demanding in terms of on-line computation and input excitation requirements.

It has been realized that there is a fundamental equivalence between parity relation and observer based designs, in that the two techniques produce identical residuals if the generators have been designed for the same specifications. A relationship, though weaker, has been found between parity relations and the parameter estimation as well, in that parity relations, designed for the isolation of parametric faults, are the minimum data-length least-squares estimators for the same. In Chapter 3, redundancy with respect to model-based methods for fault detection is discussed in detail.

3 Redundancy for Fault Detection

In this chapter different aspects of redundancies are discussed in details. There are two types of redundancies when dealing with FDI systems:

1. physical redundancy, and
2. analytical redundancy.

3.1 Physical Redundancy

Using physical redundancy, it is of common interest to classify faults into two categories: (i) faults in sensors, actuators and controllers, and (ii) faults in the plant dynamics. The types of redundancies for these two groups of faults, as well as the ways to compensate for them, are quite different. Since redundant actuators, sensors or controllers are often in hardware form, this type of redundancy is also called hardware redundancy.

In actual systems, actuators are often used to manipulate energy flow, mass flow, or to amplify the low-energy control signals to operate a process. They are the devices that mechanically drive the system. Generally speaking, the type and power levels of actuators depend on the specific applications.

Actuators for motion control may include stepper motors, DC servomotors, linear motors, or hydraulic or pneumatic motors. In process control industries, actuators can be servo-valves, solenoids and relays. Since almost all actuators require a separate power source to operate, failures in these power sources will certainly contribute to the failures of their respective actuators. Such failures are often abrupt in nature. In addition, due to their mechanical motion, actuators may also undergo wear or aging. Wear failures manifest themselves as gradual deterioration in actuator effectiveness.

Similarly, the type of sensors employed is application-dependent. Potentiometers and linear variable differential transformers are often used to measure speed and acceleration. Thermocouples, thermistors, or resistance temperature detectors are often associated with temperature measurement. Generally speaking, the signal level at a sensor output is relatively low. Additional signal conditioning and amplification often become integral parts of the sensory component. Failures in the sensors themselves or the related sensor accessories will be classified as sensor failure. The failure profile of sensors is usually different from that of actuators. The predominant problems associated with sensors are bias, loss of power, reduction in the dynamic range, or complete loss of the signal.

For a given system, redundant sensors are usually much easier to install than redundant actuators and controllers, because sensors are passive elements in the sense that they only provide the operational information of the system and do not affect the system behaviour directly unless they are in the feedback loop. There are several different strategies for introducing redundant sensors. One strategy is to use multiple dissimilar sensors at the same measurement point in the system to obtain multiple readings. These readings are then processed by a majority-voting scheme to discriminate the incorrect reading from the reliable ones. A widely accepted reliable measurement strategy in industry is the so-called Triple Modular Redundancy where three sensors are used to perform the measurement of a single quantity. Another method is to use different sensors to measure different but related system variables. Some pre-processing may be required before voting can be carried out.

In comparison with sensors, actuators are larger in size and require more power to operate. In many cases, one cannot ignore the dynamics of the actuators, which are often represented as a part of the entire system. Unlike sensors, actuators cannot be added or taken out of service easily. Therefore, the actuator redundancies can only be introduced by means of additional manipulated variables in the system. These redundant system variables may take totally different physical forms or meanings. In other words, the dynamics from each actuator to the system output may be different. Because the redundant actuators have to be in operation at all times, one cannot apply a voting scheme to the actuators, and the controller design for such systems becomes more complex.

3.2 Analytical Redundancy

It is not common to introduce system dynamic faults in terms of physical redundancy. The redundancy in this case is introduced differently, which often reflects the ability of the system to detect and diagnose component faults, and to reconfigure the control strategies to cope with them. To achieve prompt fault detection and accurate fault identification, we also have to rely on the redundant information from the system. Such information can be obtained not only from redundant sensors, but also by means of analytical relationships among related system variables. The redundancy derived from a mathematical model is often known as analytical redundancy.

The residual generation schemes of many FDI methods are based on the principle of analytical redundancies. The major hurdle in using analytical redundancy is that one has to have an accurate mathematical model representation among relevant system variables, which in practice is hard to obtain. This issue brings about a lot of research interests, such as robust FDI, and imprecise FDI for reconfigurable control. In fact, the control reconfiguration scheme can also be viewed as analytical redundancy, since different (redundant) control strategies are generated and applied to the system in the event of faults/failures. It is important to note that even though analytical redundancy is powerful in many applications, it cannot replace redundant hardware components for control purposes; this is particularly true for actuator failures.

The basic idea of analytical redundancy is the comparison of the actual behaviour of the monitored plant to the behaviour predicted on the basis of a mathematical plant model. In other words, the plant observations are checked for consistency with the mathematical model. The outcomes of the consistency checks are quantities called residuals. These residuals are nominally zero. They become nonzero as a result of faults, disturbances, noise and modelling errors. The residuals are then analyzed to arrive at a diagnostic decisions, i.e., whether there is a fault present or not, and which component is failing. Thus, any diagnostic algorithm which utilizes analytical redundancy consists of two blocks, the residual generator and the decision maker (see Fig. 3.1).

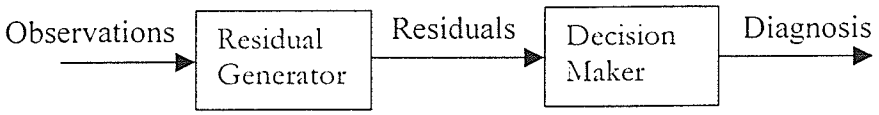


Fig. 3.1: Analytical redundancy algorithm.

While a single residual may be sufficient for fault detection, the isolation of faults requires a set of residuals. To facilitate the isolation function, the residuals are usually enhanced, that is, generated with specific isolation properties.

As discussed above, faults are performance deterioration, malfunctions or breakdowns in the monitored plant or in its instrumentation. Faults may be represented as unknown extra inputs acting on the system (i.e., additive faults) or as changes of some plant parameters (i.e., multiplicative faults). While in many cases the classification of a particular fault as additive or multiplicative follows from its nature, sometimes it may also be arbitrary. Additive faults are usually simpler to deal with, so faults should be classified as additive whenever this is reasonably justified.

A disturbance is an unknown extra input acting on the plant. Thus, there is no physical difference between a disturbance and certain additive faults; the distinction is subjective. We consider as faults those extra inputs, the presence of which we wish to detect. We consider as disturbances those which we want to ignore and by which we want to be unaffected. Some authors refer to the disturbances as 'nuisance variables'.

Additive faults and disturbances will be handled as unspecified deterministic functions of time. In general, no particular time-behaviour will be assumed or utilized in the design of residual generators. In the analysis, however, it will be useful sometimes to refer to some typical time-functions, such as drift-type, jump-type and intermittent faults and disturbances.

Modelling errors are errors or uncertainties in the parameters of the monitored system. Just like the multiplicative faults, they are discrepancies between the true system and the model, but they represent an undesirable interference with fault diagnosis. Thus, modelling errors can be considered as multiplicative disturbances. Note that, in general, it is difficult to distinguish parametric faults from certain modelling errors, though their long-term behaviour may provide some clues. Parametric faults develop over the course

of system operation, while some modelling errors may have been present from the beginning.

Table 3.1: Classification of faults, disturbances and noises.

	Additive	Multiplicative
Faults	Sensor fault Actuator fault Plant fault	Parametric Fault
Disturbances	Plant Disturbance	Modeling error
Noises	Sensor noise	

3.3 Additive Faults, Noise and Disturbances

Consider a system with multiple inputs $u(t) = [u_1(t)...u_k(t)]^T$ and multiple outputs $y(t) = [y_1(t),..., y_l(t)]^T$. For a linear discrete dynamic system, the nominal input-output relationship is

$$y(t) = M(\phi)u(t) \tag{3.1}$$

where $M(\phi)$ is the matrix that relates the output to the input. Assume that a subset $u_c(t)$ of the inputs, $u(t)$, is controlled while the rest, $u_{\bar{c}}(t)$, are measured. It is worthwhile to mention that if an input is neither controlled nor measured then it is unknown and has to be considered a disturbance. The observed variables are the command values for the controlled inputs. The measurement values are the measured inputs and the measured outputs. The following additive faults are possible:

1. Input actuator faults $\Delta u_c(t)$

2. Input sensor faults $\Delta u_M(t)$
3. Plant faults $\Delta u_P(t)$
4. Output sensor faults $\Delta y(t)$

The observed variables $u_C(t)$, $u_M(t)$ and $y(t)$ are related to the actual plant inputs $u_C^\circ(t)$ and $u_M^\circ(t)$ and the actual plant output $y^\circ(t)$, as;

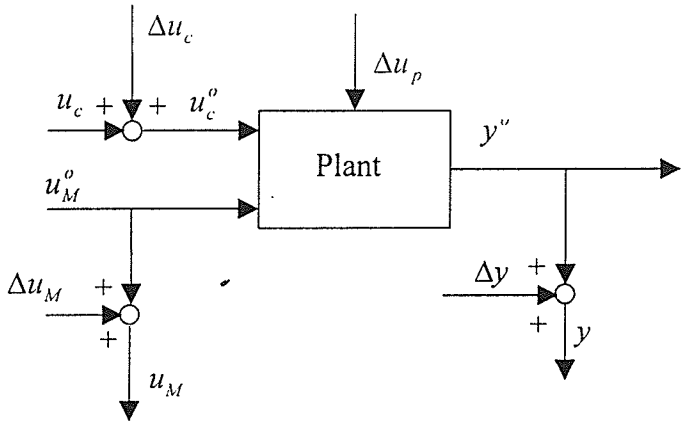


Fig. 3.2: Additive faults.

$$
\begin{aligned}
u_C^\circ(t) &= u_C(t) + \Delta u_C(t) \\
u_M^\circ(t) &= u_M(t) - \Delta u_M(t) \\
y^\circ(t) &= y(t) - \Delta y(t)
\end{aligned}
\tag{3.2}
$$

where u_C, u_M, u_P and $y(t)$ are actuator's input, sensor input, plant input and system output. $u_C^\circ, u_M^\circ, u_P^\circ$ and $y^\circ(t)$ are the corresponding true values.

The input-output relationship for the system with faults is

$$
\begin{aligned}
y(t) - \Delta y(t) &= \underbrace{[M_C(\phi) \mid M_M(\phi)]}_{M(\phi)} \begin{bmatrix} u_C(t) + \Delta u_C(t) \\ u_M(t) - \Delta u_M(t) \end{bmatrix} + S_{PF}(\phi)\Delta u_P(t) \\
&= M(\phi)u(t) + M_C(\phi)\Delta u_C(t) - M_M(\phi)\Delta u_M(t) + S_{PF}(\phi)\Delta u_P(t)
\end{aligned}
\tag{3.3}
$$

where $M_C(\phi)$ is the actuator's input-output transfer function, $M_M(\phi)$ is the sensor's input-output transfer function and $S_{PF}(\phi)$ is the plant-fault transfer function.

Defining $p(t)$ as a combined vector of additive faults:

$$p(t) = [\Delta u_C^T(t) \mid -\Delta u_M^T(t) \mid \Delta u_P^T(t) \mid \Delta y^T(t)]^T \tag{3.4}$$

Also, $S_F(\phi)$ as the combined fault transfer function:

$$S_F(\phi) = [M_C(\phi) \mid M_M(\phi) \mid S_{PF}(\phi) \mid I], \tag{3.5}$$

where I is the identity matrix. Therefore, we will have

$$y(t) = M(\phi)u(t) + S_F(\phi)p(t)$$

Now we consider the additive disturbance and noises and temporarily ignore the faults.

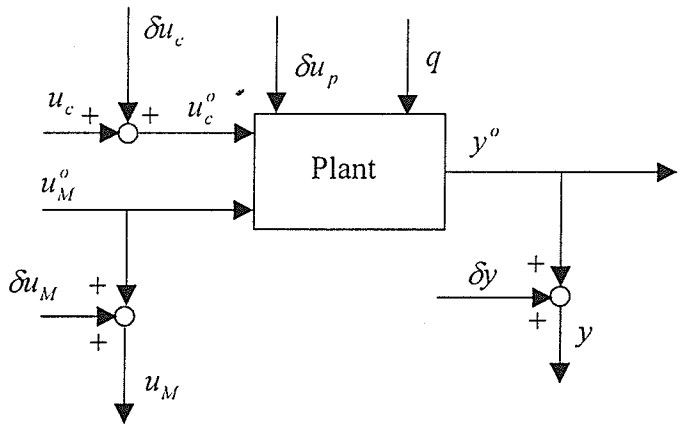


Fig. 3.3: Additive disturbances and noises.

The additive disturbances $q(t)$ act on the plant. The following noises are also possible:

1. Input actuator noise $\delta u_C(t)$
2. Input sensor noise $\delta u_M(t)$
3. Plant noise $\delta u_P(t)$
4. Output sensor noise $\delta y(t)$

Thus,

$$u_C^\circ(t) = u_C(t) + \delta u_C(t) \tag{3.6}$$

$$u_M^\circ(t) = u_M(t) - \delta u_M(t) \tag{3.7}$$

$$y^\circ(t) = y(t) - \delta y(t) \tag{3.8}$$

Similarly we can relate the input and output of the system by the following relation:

$$y(t) = M(\phi)u(t) + S_D(\phi)q(t) + S_N(\phi)v(t), \tag{3.9}$$

where $v(t)$ is the combined vector of additive noises.

$$v(t) = [\delta u_C^T(t) \mid -\delta u^T{}_M(t) \mid \delta u^T{}_P(t) \mid \delta y^T(t)]^T \tag{3.10}$$

$S_D(\phi)$ is the disturbance transfer function and $S_N(\phi)$ is defined as:

$$S_N(\phi) = [M_C(\phi) \mid M_M(\phi) \mid S_{PN}(\phi) \mid I], \tag{3.11}$$

where $S_{PN}(\phi)$ denotes the plant-noise transfer function.

Finally, if additive faults, disturbances and noise are present simultaneously, which is normally the case, then the input-output relationship becomes

$$y(t) = M(\phi)u(t) + S_F(\phi)p(t) + S_D(\phi)q(t) + S_N(\phi)v(t) \tag{3.12}$$

3.4 Multiplicative Faults and Disturbances

Consider the system modelled by Eq.(3.1). If $M^\circ(\phi)$ is the actual transfer function of the physical system, then

$$M^\circ(\phi) = M(\phi) + \Delta M(\phi), \tag{3.13}$$

where $\Delta M(\phi)$ is the discrepancy between the model and the true system. Equation (3.13) represents two conceptually different situations:

1. The discrepancy may reflect a parametric fault. In this case, it is the plant that has deviated from its earlier normal behavior, which was properly represented by the model.
2. The discrepancy may reflect a modeling error. This may be constant and present ever since the implementation of the algorithm. The error may simply be the inaccuracy of some parameters, due to perhaps identification inaccuracies, or the result of the approximation of a higher order plant with a lower order model. Another possible source of modelling error is the approximation of a nonlinear plant with a linear model. In this case the inaccuracy depends on the operating point and thus may vary with time.

We will assume, for the sake of simplicity, that the model discrepancy $\Delta M(\phi)$ is not a function of time. Though this may not be completely true, certainly the variations of the model, if any, are much slower than those of the variables.

The input-output relationship (3.1) is valid for the actual transfer function $M^{\circ}(\phi)$ which is not always entirely known:

$$y(t) = M^{\circ}(\phi)u(t) = M(\phi)u(t) + \Delta M(\phi)u(t) \tag{3.14}$$

Here the last term $\Delta M(\phi)u(t)$, is the effect of the model discrepancy. If $\Delta M(\phi)$ can be decomposed as $\Delta M_F(\phi) + \Delta M_D(\phi)$, so that the first matrix represents the parametric faults and the second the modelling errors, then (3.14) can be further written as

$$y(t) = M(\phi)u(t) + \Delta M_F(\phi)u(t) + \Delta M_D(\phi)u(t) \tag{3.15}$$

Though the model discrepancy terms appear additively, they differ from the additive fault and disturbance terms in Eq. (3.12) in an important way. The additive faults and disturbances are variables, and their coefficients in the input-output equation are time-invariant transfer functions. In contrast, the parametric faults and model errors are parameter matrices, and their coefficients in the input-output equation are variables. This is the reason why we consider the model discrepancies as multiplicative.

We now examine model discrepancies within the context of underlying parameters. One of the difficulties in handling model discrepancies is that the transfer function matrix usually has a large number of elements and each one may be accompanied by a separate scalar discrepancy. A significant simplification may be achieved if all model elements can be deduced from a small set of underlying parameters. Such underlying parameters are the parameters of a first principle characterization of the plant, and usually have direct physical meaning, such as resistance and heat transfer coefficient. Faults and modelling errors or uncertainties may first concern these underlying parameters and then propagate to the transfer function or state-space model.

Consider a set of underlying parameters $\theta = [\theta_1 ... \theta_\nu]^T$, with uncertainty $\Delta\theta = [\Delta\theta_1 ... \Delta\theta_\nu]^T$. The transfer function $M(\phi)$ is a function of the θ vector, i.e., $M(\phi,\theta)$. The underlying parameter values, which yield the model $M(\phi)$, are denoted as $\theta^{\#}$. Then $\Delta M(\phi)$ can be approximated as

$$\Delta M(\phi) = \sum_{k=1}^{\nu} M_{\theta k}(\phi)\Delta\theta_k \tag{3.16}$$

where

$$M_{\theta k}(\phi) = \frac{\partial M(\phi,\theta)}{\partial \theta_k}|_{\theta=\theta^a} \qquad k = 1....\nu \tag{3.17}$$

Note that, in general, the transfer function parameters are nonlinear functions of the underlying parameters. Therefore the accuracy of the approximation Eq. (3.16) deteriorates as the deviation from the model values grows.

With the above approximation, the model discrepancy term in Eq. (3.14) is written as

$$\Delta M(\phi)u(t) = [\sum_{k=1}^{\nu} M_{\theta k}(\phi)\Delta\theta_k]u(t) = \sum_{k=1}^{\nu}[M_{\theta k}(\phi)u(t)]\Delta\theta_k \tag{3.18}$$

or

$$\Delta M(\phi)u(t) = N(t)\Delta\theta \tag{3.19}$$

where

$$N(t) = [M\theta_1(\phi)u(t) \quad M\theta_2(\phi)u(t) \quad M\theta_3(\phi)u(t).....M\theta_\nu(\phi)u(t)]$$

and $\Delta\theta = [\Delta\theta_1, \Delta\theta_2, ...\Delta\theta_\nu]^T$

Now assume that the underlying parameter vector can be decomposed as $\theta = [\theta_F^T \mid \theta_D^T]^T$, so that discrepancies in the first group represent parametric faults while those in the second represent modelling errors. Then, with the appropriate decomposition

$$N(t) = [N_F(t) \mid N_D(t)]$$

$$\Delta M(\phi)u(t) = \Delta M_F(\phi)u(t) + \Delta M_D(\phi)u(t) = N_F(t)\Delta\theta_F + N_D(t)\Delta\theta_D \tag{3.20}$$

The above equation can be incorporated into the input-output relationship (3.12)

$$y(t) = M(\phi)u(t) + S_F(\phi)p(t) + N_F(t)\Delta\theta_F + S_D(\phi)q(t) + N_D(t)\Delta\theta_D + S_N(\phi)\nu(t) \tag{3.21}$$

This latter equation integrates the model discrepancy effects with the additive fault, disturbances and noise effects. By expressing them in terms of the underlying parameters, the model discrepancy effects become formally similar to the effects of additive faults and disturbances which act via entry matrices. However, the multiplicative faults, $\Delta\theta_F$, and disturbances, $\Delta\theta_D$, are still constants with time-varying coefficient matrices $N_F(t)$

and $N_D(t)$, while the additive faults, $P(t)$, and disturbances, $q(t)$, are variables, with time-invariant transfer function coefficients $S_F(\phi)$ and $S_D(\phi)$.

4 Fundamentals of Residual Generation

The basic idea of analytical redundancy is the comparison of the actual behaviour of the monitored plant to the behaviour predicted on the basis of a mathematical plant model. In other words, the plant observations are checked for consistency with the mathematical model. The outcomes of the consistency checks are quantities called residuals. These residuals are nominally zero. They become nonzero as a result of faults, disturbances, noise and modelling errors (Gertler, 1998). The residuals are then analyzed to arrive at a diagnostic decision. Thus, any diagnostic algorithm which utilizes analytical redundancy consists of two parts: (i) the residual generator, and (ii) the decision-maker.

In most practical situations, fault diagnosis needs to be performed in the presence of disturbances, noise and modelling errors. These interfere with the diagnosis of faults and may lead to false alarms or mis-classification of faults; therefore, the diagnostic algorithm needs to be made insensitive to the disturbances. Such techniques should be employed to suppress the effects of noise and maintain sufficient sensitivity with respect to faults.

4.1 Residual Generator

Residuals are quantities which are nominally zero. They become nonzero in response to faults, disturbances, noise and modelling error. Residuals are generated from the observation of the monitored plant. The residual generator is a linear discrete dynamic algorithm acting on the observables. Its general form is

$$r(t) = V(\phi)u(t) + W(\phi)y(t) \tag{4.1}$$

where $r(t)$ is the vector of residuals, and, $V(\phi)$ and $W(\phi)$ are transfer function matrices that relate input and output vectors to the residual vector, respectively. However the

above equation is not necessarily a residual generator. To qualify, it has to return zero residuals in the absence of no faults, disturbances, noise and modelling error i.e.,

$$V(\phi)u(t) + W(\phi)M(\phi)u(t) = 0 \tag{4.2}$$

where $y(t) = M(\phi)u(t)$ denotes the nominal input-output relationship without faults, disturbances and noise. From Eq. (4.2) we will have,

$$V(\phi) = -W(\phi)M(\phi) \tag{4.3}$$

Thus, the general residual generator can also be written as

$$r(t) = W(\phi)[y(t) - M(\phi)u(t)] \tag{4.4}$$

Equations (4.1) and (4.4) are the computational forms of the generic residual generator. Substituting term '$y(t) - M(\phi)u(t)$' from Eq. (3.21) in Eq. (4.4) one gets

$$r(t) = W(\phi)[S_F(\phi)p(t) + N_F(t)\Delta\theta_F + S_D(\phi)q(t) + N_D(t)\Delta\theta_D + S_N(t)v(t)] \tag{4.5}$$

Equation (4.5) shows how the residuals depend on the faults, disturbances and noises. Figure 4.1 shows the internal form of the residual generator.

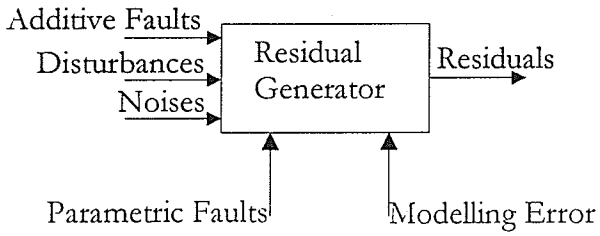


Fig. 4.1: Internal form of residual generator.

4.2 Detection Properties of the Residual Generator

Ideally, the residuals should only be affected by the faults. However, the presence of disturbances, noises and modelling errors also causes the residuals to become nonzero and thus interferes with the detection of faults. Therefore, the residual generator needs to be designed so that it is minimally affected by these nuisance inputs. Robustness is perhaps the most important requirement in residual generation. Much of the effort in designing residual generators goes into achieving sufficiently robust residual

performance. The approaches to the three classes of nuisance inputs are somewhat different. This is primarily due to the differences in their temporal behaviour.

4.2.1 Additive Disturbances

Additive disturbances are usually slow. Their temporal behaviour or frequency spectrum is similar to those of the additive faults. Therefore robust residual generation can only be achieved by explicitly decoupling the residuals from such disturbances. That is, the residual generators need to be designed so that the residuals are unaffected by the disturbances. As will be shown later, this can be done exactly if the number of disturbances is small, while only approximate decoupling is possible if this is not the case. Usually de-coupling is designed by considering the disturbances as completely unknown, that is, no particular temporal behaviour is assumed. The coefficients of additive disturbances are time-invariant transfer function matrices, therefore the residual generator may also be time-invariant. Thus, the design may be performed completely offline.

4.2.2 Multiplicative Disturbances

Multiplicative disturbances or modelling errors usually also slow. The modelling error itself is either permanently present, or arises as a result of variations of the operating point. The temporal behaviour of the modelling error effect is thus primarily determined by the coefficient of the modelling error, which is the plant input. The frequency range of the modelling error effects partially overlaps that of the faults. Thus, it is desirable to achieve explicit decoupling from the multiplicative disturbances as well. In fact, robustness in the face of modelling errors is the most significant of the robustness problems.

If the modelling errors are expressed directly in terms of the transfer function or the state-space model, the great number of affected parameters usually precludes any effective decoupling. If it is possible to represent the model uncertainty in terms of a limited number of underlying parameters, robust design may be feasible. The linear

approximation of the error propagation may somewhat compromise the accuracy of decoupling. The coefficients of the multiplicative faults are time-varying, so the residual generator designed for modelling error robustness has to be time-varying as well. We may say that the residual generator design is performed partially on-line.

4.2.3 Noise

Usually noise has zero mean and is usually in a higher frequency range than faults and disturbances. Also its statistical distribution is at least partially known or assumed. Therefore, instead of explicit decoupling, one or both of the following techniques are applied:

1. The residuals are filtered, usually by low-pass filters, to reduce the effects of noise without significantly altering those of the faults.
2. The residuals are threshold tested, instead of simply checking for nonzero values.

The threshold values may be determined by statistical considerations, on the basis of the known or assumed noise statistics. Alternatively, they may be obtained experimentally. In the latter case, they may cover not only the noise effects but also the modelling errors, which are otherwise unaccounted for. Ideally, the selection of thresholds is guided by a pre-specified trade-off between false alarms and missed detections.

The ease or difficulty of the implementation of statistical tests depends on the way that the noises propagate to the residuals. It is desirable that the noise-to-residual transfer function be of a moving average type. It is even more advantageous if the residual sequence in response to white noise input is also white.

4.2.4 Fault Sensitivity

The fault sensitivity of the residuals is an important performance characteristic of the residual generator. While other measures of sensitivity are also possible, we will introduce here the triggering limit, the value of a particular fault which brings a particular residual to its threshold, provided that no other faults and nuisance inputs are present.

Assume that k_i is the threshold for the residual r_i. The response of the same residual to a fault $p_j(t)$, with no other faults or nuisance inputs present is

$$r_i(t \mid p_j) = w_i^T(\phi) S_{Fj}(\phi) p_j(t), \tag{4.6}$$

where $w_i^T(\phi)$ is the i^{th} row of $W(\phi)$ and $S_{Fj}(\phi)$ is the j^{th} column of $S_F(\phi)$ in Eq. (4.5). Obviously, $r_i(t \mid p_j)$ is a time-function, which depends on $p_j(t)$. To be more specific, we choose a unit-step function $\varepsilon(t)$ for $p_j(t)$ and consider the response in steady-state. Thus,

$$\lim_{t\to\infty} r_i(t \mid p_j)_{step} = [w_i^T(\phi) S_{Fj}(\phi)]_{\phi=1} \tag{4.7}$$

The triggering limit η_{ij} will be

$$\eta_{ij} = \frac{k_i}{[w_i^T(\phi) S_{Fj}(\phi)]_{\phi=1}} \tag{4.8}$$

Obviously, a smaller triggering limit signifies greater fault sensitivity.

In many cases, the nominal value of the fault is known. Thus, sensitivity may be characterized by the ratio of the nominal-fault residual response to the threshold, assuming steady-state or other well-defined gain of the fault-to-residual transfer. Denoting the nominal fault sizes as p_j°, and working with the steady state gain, this ratio is

$$\xi_{ij} = \frac{p_j^\circ [w_i^T(\phi) S_{Fj}(\phi)]_{\phi=1}}{k_i} \tag{4.9}$$

It is desirable that the above ratio be slightly above one for all faults in all equations. Clearly, a ratio smaller than one signifies that the nominal fault does not bring the residual to its threshold while a ratio much larger than one indicates that even a very small fault may result in threshold crossing.

While fault sensitivities may be influenced by the filtering of the residuals, their ratio within a particular equation is fixed. Thus, the spread of the ξ_{ij} ratios within an equation is an important measure of the detection quality of that equation. This spread will be characterized by the sensitivity condition of the equation, defined as

$$\zeta_{ij} = \frac{\max_{j} \xi_{ij}}{\min_{j} \xi_{ij}} \tag{4.10}$$

4.2.5 Modelling Robustness

Modelling error robustness may be quantified in a way similar to fault sensitivity, provided the modelling error is expressed in terms of underlying parameters. Defining the limit model error as the error in a particular underlying parameter that will bring a particular residual to its threshold with no other model error or other unknown input present. The response of $r_i(t)$ to a model error $\Delta\theta_j$ is

$$r(t \mid \Delta\theta_j) = w_i^T(\phi) n_j(t) \Delta\theta_j \tag{4.11}$$

Thus the limit error is

$$\upsilon_{ij}(t) = \frac{k_i}{w_i^T(\phi) n_j(t)} \tag{4.12}$$

Higher limit error signifies lower sensitivity to model errors, that is, better robustness. Notice that the denominator in Eq. (4.12), and thus the limit error, are functions of time, i.e., it depend on the plant input.

4.3 Isolation Properties of the Residual Generator

In order to have a robust detection property, the residual generator needs to be designed to support the isolation of faults. As pointed out earlier, isolation always requires a vector of residuals. To facilitate fault isolation, the residual vector needs to have distinctive properties and unique characteristics of particular faults. Residual vectors designed with this objective in mind are referred to as enhanced residuals. There are two fundamental residual enhancement approaches: (i) directional residuals, and (ii) structured residuals. Directional residuals are designed such that, in response to a particular fault, the residual vector is confined to a fault-specific straight line. Structured residuals are the simplest

form of residuals; however they are designed such that each residual responds to a different subset of faults and is insensitive to others. When a particular fault occurs, some of the residuals respond and others do not.

With structured residuals, threshold tests are applied separately to each element of the residual vector. The outcome of the test applied to residual $r_i(t)$ is a binary variable

$$\varepsilon_i(t) = \begin{cases} 0 & if \ |r_i(t)| < k_i \\ 1 & if \ |r_i(t)| \geq k_i \end{cases} \qquad i = 1,.....,n \tag{4.13}$$

The vector $\varepsilon = [\varepsilon_1,.....,\varepsilon_n]^T$ is the fault code or signature. In this case, fault isolation amounts to comparing the actually obtained fault code to a pre-defined set of codes.

The fault codes are determined by the structure of the fault transfer matrix $W(\phi)S_F(\phi)$. A minimum requirement for structured isolation of single faults is that for each fault which is large enough to exceed its triggering limit for each residual, and with no other fault and nuisance input present, the fault code returned is different and nonzero. A residual set possessing the above property will be referred to as weakly isolating.

4.4 Computational Properties of the Generator

An important issue in the design of residual generators is the on-line computational procedure involved. On-line computations are always carried out on samples of the observable plant inputs and outputs. The residual generator usually is computationally auto regressive-moving average (ARMA), which means that the actual on-line computations are finite, but the residuals effectively reflect an infinite series of plant input and output values. It is worthwhile to design the generator to be computationally moving average (MA). According to the Eqs. (4.1) and (4.4), this requires that both $W(\phi)$ and $V(\phi) = -W(\phi)M(\phi)$ be moving average transfer functions. With such a design, the residuals are obtained from the inputs and outputs through a finite sliding window. With regards to on-line computations, dealing with multiplicative faults and disturbances is more expensive than handling additive ones. Usually decoupling techniques are less expensive than approximate ones which involve some kind of optimization at each on-line sample.

4.5 Stability of the Residual Generator

The residuals obtained from the generator need to be bounded at all times. It is reasonable to assume that the signals arising form the physical plant are bounded.

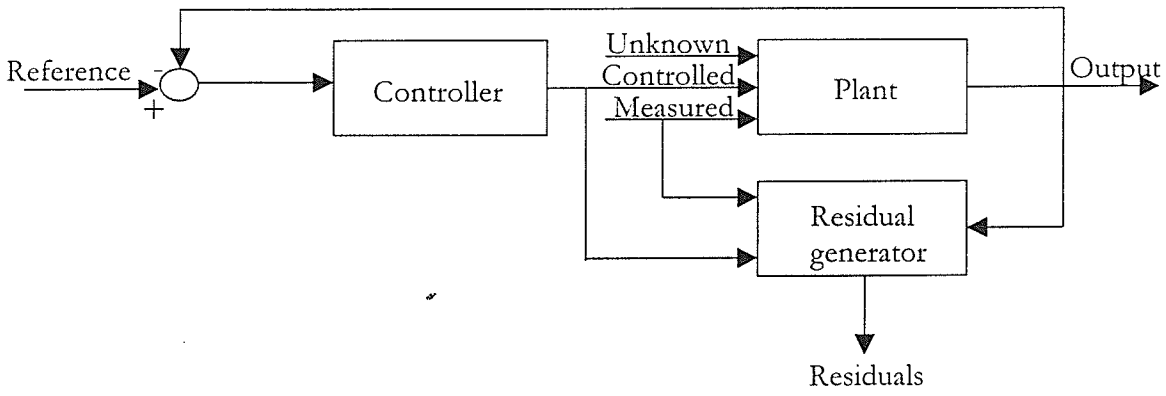


Fig. 4.2: Physical setup of the physical plant.

Thus bounded residuals require that the residual generator be bounded-input bounded-output (BIBO) stable. Figure 4.2 shows a typical set-up of the physical plant, with controller and residual generator. As far as the control system is concerned, both the measured and the unknown inputs of the plant act as disturbances. The controller is designed such that it guarantees a finite closed-loop response to both the reference input and the disturbances, even if the plant itself is unstable. The bounded signal assumption is therefore reasonable in any physically functional system.

For the generic residual generator given by Eq. (4.4), the BIBO stability requirement has to be satisfied for both $W(\phi)$ and $W(\phi)M(\phi)$. If the generator is computationally moving average, BIBO stability follows naturally. For ARMA generators, $W(\phi)$ has to be so chosen that it is stable itself and stabilizes $W(\phi)M(\phi)$. The latter requires the cancellation of any unstable poles $M(\phi)$ may have. This can be done without the usual risks that pole-zero cancellation carries in controller design because the residual generator is not in the loop and because it is designed with the plant model $M(\phi)$, which

generator is not in the loop and because it is designed with the plant model $M(\phi)$, which is known exactly, in contrast to the real plant transfer function $M^{\circ}(\phi)$ which is not known.

4.6 Observer-Based Residual Generation

The overall procedure of fault detection by state estimation (observer or Kalman filter) consists of the following two steps:

1. generation of residuals by state estimation, and
2. evaluation of the residuals by a decision logic.

The purpose of residual generation via state estimation is to reconstruct the outputs of the process with the aid of observers or Kalman filters. Usually fault detection and isolation techniques incorporating residual generation are classified as either:

1. parity equation residual generation, or
2. observer-based residual generation.

We will deal with observer-based residual generation methods, as this is the primary objective of this work. Observer-based residual generation methods can be further classified into linear observer-based residual generation and nonlinear observer-based residual generation.

4.6.1 Linear Observer-Based Residual Generation

There are different approaches toward the design of a linear observer for residual generation based on different aspects of needs and system dynamics. The Unknown Input Observer, Eigenstructure Assignment and Detection Filter schemes are some well-known methods of residual generation. The unknown input observer design as proposed by Viswanadham and Srichander (1987) and Hou and Muller (1992), consists of transforming the system equations so that the state observer can be divided into two parts: one part that can be directly obtained from the measurements, and another part consisting of the states that have to be estimated. A reduced-order observer can be designed to

estimate these states, and the observer gains are so selected that they decouple the observer dynamics from the unknown inputs. Hou and Muller (1991) extended the unknown input observer (UIO) to systems whereby the unknown input enters the measurement equation. The unknown input method of residual generation can also be implemented using the generalized observer scheme as described by Frank et al. (Patton et al., 1989).

Patton and Chen (1991) demonstrated the use of eigenstructure assignment in FDI. Both left and right eigenvector assignment may be used to produce robust residuals that maintain a fixed direction in the output error space. Yuksel et al. (1971) proposed algorithms for linear observers for index-invariant uniformly observable time-varying linear finite-dimensional multivariable systems. The results obtained indicated that asymptotic estimators can be employed in optimally designed regulators.

The detection filter (DF) approach to the residual generation problem was first proposed by Beard (1971) and Jones (1973). In this scheme, an observer is designed such that in the presence of faults the residual vector in the output space lies in a well-defined direction that corresponds to and allows for the identification of the fault that has occurred. This approach was later extended by many researchers. For example, Massoumnia (1986) reformulated the detection filter problem using a geometric approach. White and Speyer (1987) adopted a spectral approach. The approach considers the eigenvalue and eigenvector problem directly, without requiring the selection of a generator vector as in Beard's method. The main limitation of this method, however, is that the eigenvalues and the eigenvectors of both the detection and the completion spaces must be specified simultaneously.

4.6.2 Nonlinear Observer-Based Residual Generation Methods

The approach adopted in this thesis is based on a nonlinear observer. The methods discussed in the previous section were for linear systems, which may not work properly for systems having nonlinearities. Hence, some authors such as Hengy and Frank (1986) and Seliger and Frank (1991) have proposed residual generation schemes that use the theory of nonlinear observers. Seliger and Frank (1991) proposed a nonlinear unknown

disturbance or unknown inputs. Hengy and Frank (1986) presented a scheme for detecting and isolating faults in a specific component of a complex system using a nonlinear observer. Yu et al. (1994) proposed a nonlinear observer for a bilinear system with unknown inputs. In this thesis, the approach by Rajamani (1998) is adopted and will be discussed in detail.

Rajamani (1998) developed an observer for a Lipschitz nonlinear system described by

$$\left.\begin{aligned} \dot{x} &= Ax + \Phi(x,u) \\ y &= Cx \end{aligned}\right\} \tag{4.14}$$

where $\Phi(x,u)$ is a Lipschitz nonlinearity with a Lipschitz constant γ, i.e.,

$$\|\Phi(x,u) - \Phi(\hat{x},u)\| \le \gamma \|x - \hat{x}\| \tag{4.15}$$

The observer is assumed to be of the following form:

$$\dot{z} = Az + \Phi(z,u) + L[y - Cz] \tag{4.16}$$

The estimation error dynamics is given by

$$\dot{e} = \dot{x} - \dot{z} = (A - LC)e + [\Phi(x,u) - \Phi(z,u)] \tag{4.17}$$

If

$$\gamma < \frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)} \tag{4.18}$$

where $(A - LC)^T P + P(A - LC) = -Q$, then Eq. (4.16) yields an asymptotically stable estimator for the system in Eq. (4.14).

Rajamani and Hedrick (1995) further extended the above nonlinear observer. They considered the class of nonlinear dynamical systems described by

$$\left.\begin{aligned} \dot{x} &= Ax + \Phi(x,u) + bf(x,u)\theta \\ y &= Cx \end{aligned}\right\} \tag{4.19}$$

where $x \in R^n$, $y \in R^m$, $\theta \in R^r$, $f: R^n \to R^{s\times r}\Phi: R^n \to R^n$, $b \in R^{n\times s}$, $C \in R^{m\times n}$.

If

1) There exists a positive definite symmetric matrix P such that

$$b^T P = C_1 \tag{4.20}$$

2) $\Phi(..)$ and $f(..)$ are Lipschitz in x with Lipschitz constants γ_1 and γ_2, respectively, for all $x, \hat{x} \in R^n$, i.e.,

$$\|\Phi(x,u) - \Phi(\hat{x},u)\| \le \gamma_1 \|x - \hat{x}\|$$

$$\|f(x,u) - f(\hat{x},u)\| \le \gamma_2 \|x - \hat{x}\|$$

3) The vector of unknown parameters θ is bounded in the following sense

$$\|\theta\| \le \gamma_3 \tag{4.21}$$

4) A gain matrix L can be chosen such that

$$\gamma_1 + \gamma_1\gamma_2\|b\| < \frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)} \tag{4.22}$$

where Q is a positive definite symmetric matrix satisfying the Lyapunov equation

$$(A - LC)^T P + P(A - LC) = -Q \tag{4.23}$$

Then, the following nonlinear observer can be proposed:

$$\dot{z} = Az + \Phi(z,u) + bf(z,u)\hat{\theta} + L[y - Cz] \tag{4.24}$$

$$\dot{\hat{\theta}} = \frac{1}{\psi} f(z,u)^T [y - Cz] \qquad (\psi > 0) \tag{4.25}$$

5 Statistical Methods for Fault Detection

Techniques for fault detection that take into account the existence of error in the process variables and coefficients can be implemented separately or in conjunction with techniques which ignore the error. To make correct decisions in the face of uncertainty, the analyst must be able to choose rationally from among alternatives. Hence, as random errors exist in nearly all process measurements, sound decision-making requires a variety of skills on the part of the analyst. The objective of the analysis may be to test a hypothesis, to develop a suitable relationship among variables, or to discriminate among possible faults. However, no matter what the objective of the measurements and subsequent analysis is, the tools of analysis, to a large extent, make use of the discipline of statistics.

5.1 Hypothesis Testing

In hypothesis testing, one tests a hypothesis, H_o, against one or more alternate hypotheses (H_1, H_2,...) that are spelled out or implied. For example, the hypothesis H_o might be that $\mu = 10$; two alternate hypotheses might be $H_1 : \mu > 10$, and $H_2 : \mu < 10$. Similarly, the hypothesis to be tested might be that there is no fault in a process as compared with the alternate hypothesis that there is a fault. Suppose we know the probability density function, $p(\hat{\theta})$, for an estimate $\hat{\theta}$, which is an unbiased estimate of θ. We assume that the representation of the random variable $\hat{\theta}$ by $p(\hat{\theta})$ is correct, and that the ensemble value of θ is θ_o. We ask the following question: if we presume as true the hypothesis that $\theta = \theta_o$, by how much must $\hat{\theta}$ differ from θ_o before we reject the hypothesis because it seems to be wrong?

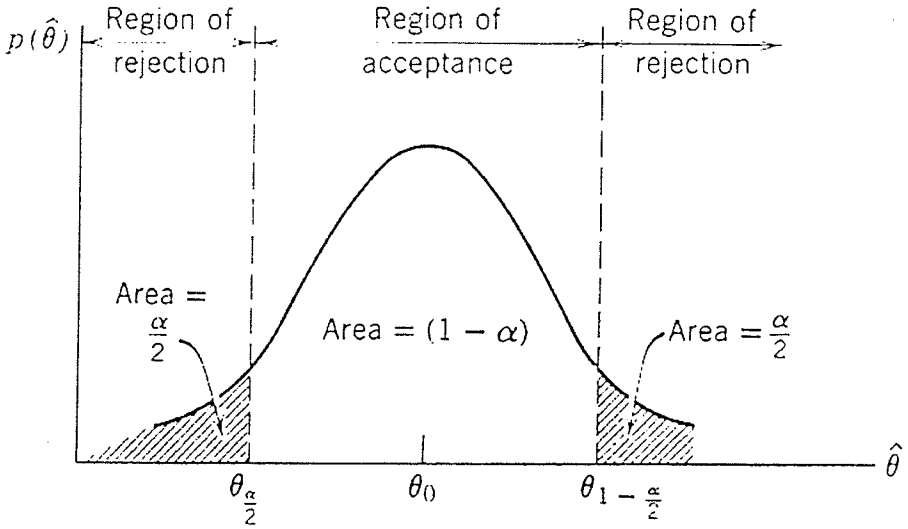


Fig. 5.1: Regions of rejection and acceptance for a symmetric hypothesis test (Himmelblau, 1978; p.61).

Fig. 5.1 helps to answer the question. If the hypothesis $\theta = \theta_o$ is true, $E\{\hat{\theta}\} = \theta_o$ as is shown in the figure. The probability that the value of $\hat{\theta}$ would be equal to or less than $\theta_{\frac{\alpha}{2}}$ is

$$P\{\hat{\theta} \le \theta_{\frac{\alpha}{2}}\} = \int_{-\infty}^{\theta_{\frac{\alpha}{2}}} p(\hat{\theta})d\hat{\theta} = \frac{\alpha}{2} \tag{5.1}$$

and, because of symmetry

$$P\{\hat{\theta} > \theta_{1-\frac{\alpha}{2}}\} = \int_{\theta_{1-\frac{\alpha}{2}}}^{\infty} p(\hat{\theta})d\hat{\theta} = \frac{\alpha}{2} \tag{5.2}$$

To reach a decision concerning the hypothesis, we select a value of α, which is termed the level of significance for the test, before collecting the sample. α is selected to be small enough so that the user regards it quite improbable that $\hat{\theta}$ will exceed the selected value of $\theta_{1-\frac{\alpha}{2}}$ or be less than $\theta_{\frac{\alpha}{2}}$. For example, α might be 0.04 or less. Once the sample is collected and $\hat{\theta}$ is calculated, if $\hat{\theta}$ is larger than $\theta_{1-\frac{\alpha}{2}}$ or smaller than $\theta_{\frac{\alpha}{2}}$, the hypothesis is rejected. Otherwise, it is accepted. The range of values of $\hat{\theta}$ for which the

hypothesis is rejected is called the region of rejection, while the range of $\hat{\theta}$ for which the hypothesis is accepted is called the region of acceptance.

The test described above is a two-sided test. A one-sided test can be based on either $\hat{\theta}$ being greater than some $\theta_{1-\alpha}$, with the hypothesis $\theta = \theta_o$ being rejected if $\hat{\theta}$ is greater than $\theta_{1-\alpha}$, or on $\hat{\theta}$ being less than θ_α. Rejecting the hypothesis does not mean immediately sounding an alarm, but instead calls for a careful examination of the experimental procedure and data to ascertain if anything went wrong with the collection of measurements or the instrumentation. Investigation into the causes of defects in the method of procedure can be most rewarding.

The simplest structure for testing is to imagine that a dichotomy of states exist for a random variable:

1. H_o: x is the true state of the random variable (the null hypothesis).
2. H_1: x is not the true state of the variable (the alternate hypothesis).

For example, hypothesis H_o states that the ensemble mean of a process variable has not changed, while H_1 states that the process mean has changed.

In hypothesis testing, a decision is made as follows. Based on the assumption that the null hypothesis is true, if the statistic calculated from the random experimental sample falls outside the region of acceptance, the null hypothesis, H_o, is rejected and H_1 is accepted. Otherwise, H_o is accepted and H_1 rejected.

Two types of errors can be distinguished in testing a hypothesis:

1. An error of the first kind which is caused by rejecting the hypothesis when it is true, causing a false alarm.
2. An error of the second kind which is caused by not rejecting the hypothesis when it is false, i.e., not causing an alarm when a fault exists. It is normally called a missing alarm.

Which probability is to be optimized depends on the process, its instrumentation and the costs of each decision. Thus the probabilities of any process being monitored can be classified in four ways:

1. the probability of calling a good process good,

2. the probability of calling a good process faulty,
3. the probability of calling a faulty process faulty, and
4. the probability of calling a faulty process good,

Sequential probability ratio tests are very powerful hypothesis tests in statistics. One of the sequential probability ratio tests, sequential test of Wald, is discussed in the next section.

5.2 Sequential Test of Wald

Sequential analysis is a method of statistical inference whose characteristic feature is that the number of observations required by the procedure is not determined in advance of the experiment. The decision to terminate the experiment depends at each stage on the results of the observations made previously (Wald, 1947). A merit of the sequential method, as applied to testing statistical hypotheses, is that test procedures can be constructed which require, on the average, a substantially smaller number of observations than equally reliable test procedures based on a predetermined number of observations. The sequential probability ratio test frequently results in a saving of about 50% in the number of observations over the most efficient test procedure based on a fixed number of observations.

In the theory of testing hypotheses, the number of observations, i.e., the size of the sample on which the test is based, is treated as a constant for any particular problem. An essential feature of the sequential test, as distinguished from the current test procedure, is that the number of observations required by the sequential test depends on the outcome of the observations and is therefore not a predetermined variable but a random one.

In the sequential method of testing a hypothesis H may be described as follows. A rule is given for making one of the following three decisions at any stage of the experiment:

1. To accept the hypothesis H,
2. To reject the hypothesis H
3. To continue the experiment by making an additional observation.

Thus, such a test procedure is carried out sequentially. On the basis of the first observation one of the above three decisions is made. If the first or second decision is made, the process is terminated. If the third decision is made, a second trial is performed. Again, on the basis of the first two observations one of the three decisions is made. If the third decision is made, a third trial is performed, and so on. The process is continued until either the first or the second decision is made. The number of observations, n, required by such a test procedure is a random variable, since the value of n depends on the outcome of the observations.

For each positive integer value m, we shall denote by M_m the totality of all possible samples $(x_1,...,x_m)$ of size m. We shall also refer to M_m as an m-dimensional sample space. A rule for making one of the three decisions at any stage of the experiment can be described as follows. The m-dimensional sample space is split into three mutually exclusive parts: $R_m^{\,0}$, $R_m^{\,1}$ and R_m. After the first observation x_1 has been drawn, the hypothesis H that is being tested is accepted if x_1 lies in $R_1^{\,0}$. H is rejected if x_1 lies in $R_1^{\,1}$ or a second observation is made if x_1 lies in R_1. If the third decision is made and a second observation x_2 drawn, H is accepted, rejected, or a third observation is drawn, depending whether the observed sample (x_1,x_2) lies in $R_2^{\,0}$, $R_2^{\,1}$, or R_2. If (x_1,x_2) lies in R_2 a third observation x_3 is drawn and one of the three decisions is made according to where (x_1,x_2,x_3) lies. As can be seen a sequential test is completely defined by defining the sets, $R_m^{\,0}$, $R_m^{\,1}$, and R_m for all positive integer values m. Since $R_m^{\,0}$, $R_m^{\,1}$, and R_m are mutually exclusive and add up to the whole sample space M_m, it is sufficient to define any two of the sets, $R_m^{\,0}$, $R_m^{\,1}$, and R_m. Any one of the three sets, $R_m^{\,0}$, $R_m^{\,1}$, and R_m consists precisely of all those samples which are not contained in the other two.

We shall call a sample $(x_1,...,x_m)$ ineffective if it contains an initial segment $(x_1,...,x_{m'})$, where $m' < m$, such that $(x_1,...,x_{m'})$ lies in $R_{m'}^{\,0}$, or in $R_{m'}^{\,1}$. A sample which is not ineffective will be said to be an effective sample. Clearly, for a sequential test procedure we shall have an effective sample at any stage of the experiment. Thus, in

defining sets $R_m^{\,0}$, $R_m^{\,1}$ and R_m, we may disregard ineffective samples. In other words, it is sufficient to state in which of the sets $R_m^{\,0}$, $R_m^{\,1}$, and R_m each effective sample $(x_1,...,x_m)$ should be included, since ineffective samples cannot occur during the sequential process.

The requirements regarding the tolerated risks are satisfied by the sequential probability ratio test of strength (α,β) for testing the hypothesis that $\theta=\theta_o$ against the alternative that $\theta=\theta_1$. This sequential test is described below.

Let $x_1,x_2,....,x_m$ be the successive observations on x. The probability density of the sample, $x_1,x_2,....$ and, x_m is given by

$$p_{om}=\frac{1}{(2\pi)^{\frac{m}{2}}\sigma^m}e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(x_k-\theta_o)^2} \quad \text{if } \theta=\theta_o \tag{5.3}$$

and by

$$p_{1m}=\frac{1}{(2\pi)^{\frac{m}{2}}\sigma^m}e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(x_k-\theta_1)^2} \quad \text{if } \theta=\theta_1 \tag{5.4}$$

The probability ratio $\frac{p_{1m}}{p_{om}}$ is computed at each stage of the inspection. Additional observations are taken as long as

$$B<\frac{p_{1m}}{p_{om}}=\frac{e^{-\frac{1}{2\sigma^2}\sum(x_k-\theta_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum(x_k-\theta_o)^2}}<A \tag{5.5}$$

Inspection is terminated with acceptance if

$$\frac{p_{1m}}{p_{om}}=\frac{e^{-\frac{1}{2\sigma^2}\sum(x_k-\theta_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum(x_k-\theta_o)^2}}\le B \tag{5.6}$$

Inspection is terminated with the rejection if

$$\frac{p_{1m}}{p_{om}} = \frac{e^{-\frac{1}{2\sigma^2}\sum(x_k-\theta_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum(x_k-\theta_o)^2}} \geq A \tag{5.7}$$

Approximate values of A and B are given by $\frac{(1-\beta)}{\alpha}$ and $\frac{\beta}{(1-\alpha)}$, respectively;

By taking the logarithms and simplifying inequalities (5.5), (5.6) and (5.7), the following relations will hold:

$$\ln\frac{\beta}{1-\alpha} < \frac{\theta_1-\theta_o}{\sigma^2}\sum_{k=1}^{m}x_k + \frac{m}{2\sigma^2}(\theta_0^2-\theta_1^2) < \ln\frac{1-\beta}{\alpha}, \tag{5.8}$$

$$\frac{\theta_1-\theta_o}{\sigma^2}\sum_{k=1}^{m}x_k + \frac{m}{2\sigma^2}(\theta_0^2-\theta_1^2) \leq \ln\frac{\beta}{1-\alpha}, \tag{5.9}$$

and

$$\frac{\theta_1-\theta_o}{\sigma^2}\sum_{k=1}^{m}x_k + \frac{m}{2\sigma^2}(\theta_0^2-\theta_1^2) \geq \ln\frac{1-\beta}{\alpha} \tag{5.10}$$

Further simplification in carrying out the test procedure can be achieved by adding $(-\frac{m}{2\sigma^2})(\theta_o^2-\theta_1^2)$ to both sides of the inequalities (5.8), (5.9), and (5.10) and then dividing these inequalities by $\frac{(\theta_1-\theta_o)}{\sigma^2}$. These operations transform the inequalities (5.8), (5.9), and (5.10) into

$$\frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\frac{\theta_o+\theta_1}{2} < \sum_{k=1}^{m}x_k < \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{1-\beta}{\alpha} + m\frac{\theta_o+\theta_1}{2}, \tag{5.11}$$

$$\sum_{k=1}^{m}x_k \leq \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\frac{\theta_o+\theta_1}{2}, \tag{5.12}$$

and

$$\sum_{k=1}^{m}x_k \geq \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{1-\beta}{\alpha} + m\frac{\theta_o+\theta_1}{2} \tag{5.13}$$

Using the inequalities (5.11), (5.12), and (5.13), the inspection plan may be carried out as follows. For each m, compute the acceptance number

$$a_m = \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\frac{\theta_o+\theta_1}{2} \tag{5.14}$$

$$r_m = \frac{\sigma^2}{\theta_1 - \theta_o} \ln \frac{1-\beta}{\alpha} + m \frac{\theta_o + \theta_1}{2} \tag{5.15}$$

These acceptance and rejection numbers are best computed before inspection starts. Inspection is continued as long as $a_m < \sum x_k < r_m$. At the first instance when $\sum x_k$ does not lie between a_m and r_m, inspection is terminated. The lot is accepted if $\sum x_k \le a_m$, and the lot is rejected if $\sum x_k \ge r_m$. As an example, consider the test procedure carried out graphically as shown in Fig. 5.2. In this example $\theta_o = 135$, $\theta = 150, \alpha = 0.01$, $\beta = 0.03$ and $\sigma = 25$. The sampling inspection is terminated at $m = 20$ with the acceptance of the lot.

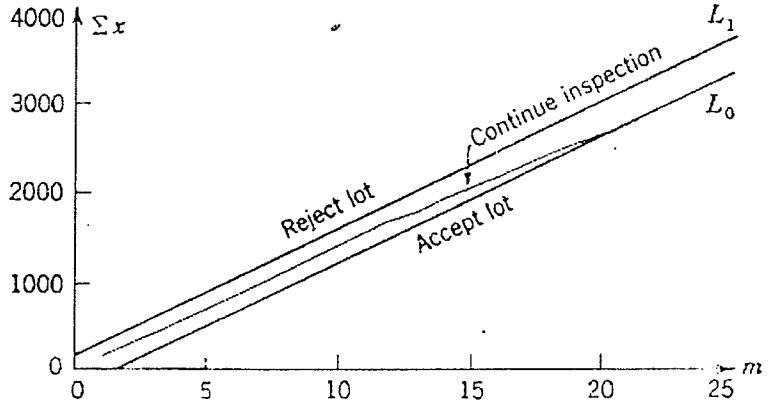


Fig. 5.2: Graphical representation of Wald's test (Wald, 1947; p.120).

The number of observations, m, is measured along the horizontal axis. The points (m, a_m) will lie on a straight line L_o and the points (m, r_m) will lie on a parallel line L_1. We draw the parallel lines L_o and L_1 before inspection starts. The points $(m, \sum_{k=1}^{m} x_k)$ are plotted as inspection goes on. Inspection is continued as long as the plotted points $(m, \sum_{k=1}^{m} x_k)$ lie between the lines L_o and L_1. Inspection is terminated at the first time when the point $(m, \sum_{k=1}^{m} x_k)$ does not lie between L_o and L_1. If it lies on L_o or below the lot is accepted, and if it lies on L_1 or above the lot is rejected.

The common slope of the lines L_o and L_1 is given by

$$s = \frac{\theta_o + \theta_1}{2} \tag{5.16}$$

The intercept of L_o with y-axis is equal to

$$h_o = \frac{\sigma^2}{\theta_1 - \theta_o} \ln \frac{\beta}{1-\alpha} \tag{5.17}$$

and the intercept of L_1 with the y-axis is given by

$$h_1 = \frac{\sigma^2}{\theta_1 - \theta_o} \ln \frac{1-\beta}{\alpha} \tag{5.18}$$

6 Experimental Setup and Mathematical Modelling

6.1 Experimental Hydraulic Test Station

Figure 6.1 shows the experimental test rig on which all the experiments were performed.

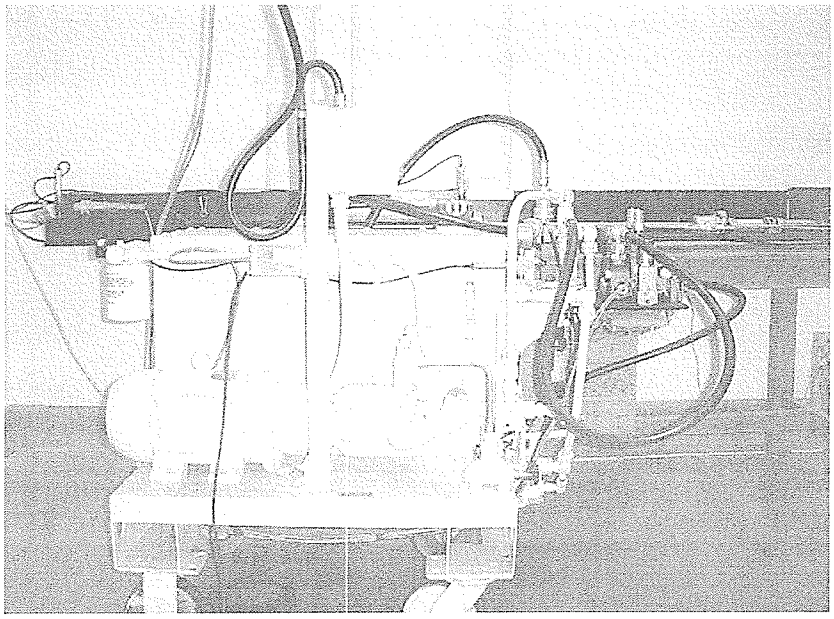


Fig. 6.1: Experimental test rig.

The system consists of a hydraulic pressure source and electro-hydraulic proportional valves connected to a hydraulic cylinder by flexible hoses. The system can switch between two different valves enabling to work in both constant flow and constant pressure operational modes.

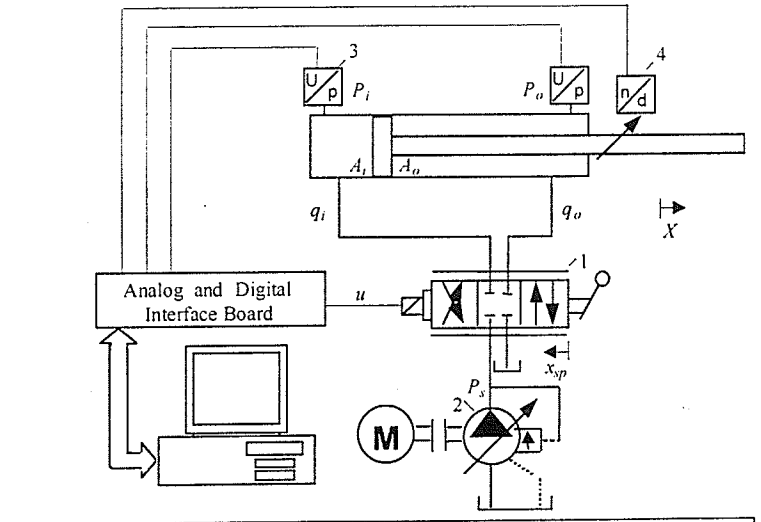
An asynchronous 3700 Watt electrical motor drives the variable displacement pump. The output flow of the pump can be set to a maximum flow rating of 28 litres per minute at a nominal speed of 1800 rpm. The pump pressure can be continuously regulated up to

250 bar (3625 psi). The system works in either constant pressure or load sensing mode. It can also be set to simply provide a constant flow regardless of the operating pressure.

The close-centre valve is a proportional valve with load sensing capability (Danfoss PVG/PVEH model). The positioning of the valve spool is based on the pulse width modulation principle. The reaction time of the valve from the neutral position to maximal spool travel is rated at 120 msec.

The piston and rod-side areas of the actuator are $0.00114m^2$ and $0.000633m^2$, respectively. The actuator has a maximum stroke length of about 1m. For monitoring of the supply pressure and the pressures at the both ends of the cylinder, three pressure transducers are installed and connected to the data acquisition board. The piston position is obtained by an incremental encoder with a resolution of approximately 0.066mm. A 486 DX2/66 personal computer is used for sampling, which also performs all control actions. Using this one-axis model of the hydraulic manipulator, one can focus on problems which are exclusively related to the control of typical industrial hydraulic actuators.

6.2 System Modelling



1 Proportional valve	3 Pressure transducer
2 Pump with pressure regulator	4 Incremental encoder

Fig. 6.2: Schematic diagram of hydraulic actuator.

The governing nonlinear equations describing the fluid flow distribution in the valve, shown above in Fig. 6.2, are written in their simplest forms as follows (Merrit, 1967):

$$Q_i = \begin{cases} k_d w x_{sp} \sqrt{P_s - P_i} & \text{If } x_{sp} \geq 0 \text{ (extension)} \\ k_d w x_{sp} \sqrt{P_i - P_r} & \text{If } x_{sp} < 0 \text{ (retraction)} \end{cases} \tag{6.1}$$

$$Q_o = \begin{cases} k_d w x_{sp} \sqrt{P_o - P_r} & \text{If } x_{sp} \geq 0 \text{ (extension)} \\ k_d w x_{sp} \sqrt{P_s - P_o} & \text{If } x_{sp} < 0 \text{ (retraction)} \end{cases} \tag{6.2}$$

where Q_i and Q_o represent fluid flows into and out of the valve, respectively, k_d is the metering coefficient, w is the orifice area gradient that relates the spool displacement (x_{sp}) to the orifice area, P_s is the pump pressure, P_i and P_o are the line input and output pressures, respectively, and P_r is the return pressure.

Continuity equations for oil flow through the cylinder, neglecting the leakage flow across the actuator's piston are:

$$Q_i = \frac{V_i(x)}{\beta} \dot{P}_i + A_i \dot{x} \tag{6.3}$$

$$Q_o = -\frac{V_o(x)}{\beta} \dot{P}_o + A_o \dot{x} \tag{6.4}$$

where A_i and A_o are the piston effective areas. β is the effective bulk modulus, and V_i and V_o are the volumes of fluid trapped at the sides of actuator. They can be expressed as functions of actuator linear displacement.

$$V_i(x) = \overline{V}_i + xA_i \tag{6.5}$$

$$V_o(x) = \overline{V}_o - xA_o \tag{6.6}$$

where $\overline{V}_i$ and $\overline{V}_o$, are the initial volumes trapped in the blind and rod sides of the actuator. Continuity equations for oil flow through the cylinder, considering the leakage flow across the actuator's piston are:

$$Q_i - C_i(P_i - P_o) = \frac{V_i(x)}{\beta} \dot{P}_i + A_i \dot{x} \tag{6.7}$$

$$C_i(P_i - P_o) - C_e P_o - Q_o = -\frac{V_o(x)}{\beta}\dot{P}_o + A_o\dot{x} \tag{6.8}$$

where C_i is the cross-port leakage coefficient of piston and C_o is the external leakage coefficient of piston. The relationship between the spool displacement, x_{sp}, and the input voltage, u, to the proportional valve can be expressed as a first-order differential equation.

$$u = \frac{\tau}{k_{sp}}\dot{x}_{sp} + \frac{1}{k_{sp}}x_{sp} \tag{6.9}$$

where τ and k_{sp} are gains describing the valve dynamics.

Rigid body dynamic equations are:

$$F = (P_i A_i - P_o A_o) = m\ddot{x} + f_d\dot{x} \tag{6.10}$$

where m is the load. Defining the following state vectors

$$x = [x_1, x_2, x_3, x_4]^T = [v, p_i, p_o, x_{sp}]^T$$

and rearranging the differential equations as (for $x_{sp} \geq 0$);

$$\dot{x}_1 = v = \frac{1}{m}[-f_d x_1 + A_i x_2 - A_o x_3] \tag{6.11}$$

$$\dot{x}_2 = \dot{P}_i = \frac{\beta}{V_i}[(k_d \omega x_4 \sqrt{p_s - x_2}) - A_i x_1] \tag{6.12}$$

$$\dot{x}_3 = \dot{P}_o = \frac{\beta}{V_o}[A_o x_1 - k_d \omega x_4 \sqrt{x_3 - p_r}] \tag{6.13}$$

$$\dot{x}_4 = \dot{x}_{sp} = \frac{1}{\tau}[u k_{sp} - x_4] \tag{6.14}$$

$$\dot{x} = \begin{Bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{Bmatrix} = \begin{bmatrix} \frac{-f_d}{m} & \frac{A_i}{m} & \frac{-A_o}{m} & 0 \\ \frac{-\beta A_i}{V_i} & 0 & 0 & \frac{\beta}{V_i}[k_d \omega \sqrt{p_s - x_2}] \\ \frac{\beta A_o}{V_o} & 0 & 0 & \frac{-\beta}{V_o}[k_d \omega \sqrt{x_3 - p_r}] \\ 0 & 0 & 0 & \frac{-1}{\tau} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{k_{sp}}{\tau} \end{bmatrix} u \tag{6.15}$$

7 Development of FDI Scheme for Hydraulic System

7.1 Design of Nonlinear Observer

Assume that the system state $x=[x_1,x_2,x_3,x_4]^T$ of the dynamic model is to be estimated by the state $z=[z_1,z_2,z_3,z_4]^T$. The state observer is

$$\dot{z}=\begin{bmatrix} \frac{-f_d}{m} & \frac{A_i}{m} & \frac{-A_o}{m} & 0 \\ \frac{-\beta A_i}{V_i} & 0 & 0 & \frac{\beta}{V_i}[k_d\omega\sqrt{p_s-x_2}] \\ \frac{\beta A_o}{V_o} & 0 & 0 & \frac{-\beta}{V_o}[k_d\omega\sqrt{x_3-p_r}] \\ 0 & 0 & 0 & \frac{-1}{\tau} \end{bmatrix}\begin{Bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{Bmatrix}+\begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{k_{sp}}{\tau} \end{bmatrix}u+\begin{bmatrix} K_1(x_1-z_1) \\ K_2(x_1-z_1) \\ K_3(x_1-z_1) \\ K_4(x_1-z_1) \end{bmatrix} \tag{7.1}$$

where K_1,K_2,K_3,K_4 are gains of the observer, $x_1=v$ is the output of the actual system and $z_1=\hat{v}$ is the output of the observer.

7.2 Stability Verification

7.2.1 Basic Theory

In this section some general and useful conditions are derived which guarantee the stability for the class of observers described by the Eq. (7.1). Consider a system described as

$$\dot{x}=\phi(x,u)=Ax+f(x,u) \tag{7.2}$$

$$y=Cx$$

The differential equation of the observer for the system shown above is presented as

$$\dot{z} = Az + f(z,u) + K(y - Cz) \tag{7.3}$$

Function $f(..)$ describes the dynamic model of the system as well as the observer. It is continuous in x and z, and satisfies the following Lipschitz condition:

$$\|f(x,u) - f(\hat{x},u)\| \le \gamma \|x - \hat{x}\| \qquad x, \hat{x} \in R^n \tag{7.4}$$

where γ is a constant. The proof of the existence of γ to satisfy condition (7.4) and the formulation of the observer gain K is important because of the nonlinearity characteristic of the observer model defined by the function f. It was shown by Rajamani and Hedrick (1995) that if the observer gain matrix K is chosen such that:

$$\gamma < \frac{\lambda_{\min}(q)}{2\lambda_{\max}(p)} \tag{7.5}$$

where λ denotes the eigenvalues of matrices p and q that are positive definite, symmetric, and satisfy the Lyapunov equation:

$$(A - KC)^T p + p(A - KC) = -q \tag{7.6}$$

then the observer defined by Eq. (7.3) is stable.

7.2.2 Stability Proof

Comparing Eqs. (6.15) and (7.2), one finds the following relations:

$$A = \begin{bmatrix} -\frac{f_d}{m} & \frac{A_i}{m} & -\frac{A_o}{m} & 0 \\ -\frac{\beta A_i}{V_i} & 0 & 0 & 0 \\ \frac{\beta A_o}{V_o} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\tau} \end{bmatrix} \tag{7.7}$$

$$f(x,u)=\begin{Bmatrix} 0 \\ [a_1\sqrt{P_s-x_2}]x_4 \\ [-a_2\sqrt{x_3-p_r}]x_4 \\ \frac{k_{sp}}{\tau}u \end{Bmatrix} \tag{7.8}$$

where $a_1=\frac{\beta}{V_i}k_d\omega,\ a_2=\frac{\beta}{V_o}k_d\omega$

f(x,u) is Lipschitz nonlinear with a Lipschitz constant γ if and only if

$$\|f(x,u)-f(\hat{x},u)\|\le\gamma\|x-\hat{x}\| \tag{7.9}$$

$$\left\|\begin{Bmatrix} 0 \\ [a_1\sqrt{p_s-x_2}]x_4 \\ [-a_2\sqrt{x_3-p_r}]x_4 \\ \frac{k_{sp}}{\tau}u \end{Bmatrix}-\begin{Bmatrix} 0 \\ [a_1\sqrt{p_s-\hat{x}_2}]\hat{x}_4 \\ [-a_2\sqrt{\hat{x}_3-p_r}]\hat{x}_4 \\ \frac{k_{sp}}{\tau}u \end{Bmatrix}\right\|\le\gamma\left\|\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix}-\begin{Bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{Bmatrix}\right\| \tag{7.10}$$

$$\sqrt{[(a_1\sqrt{p_s-x_2})x_4-(a_1\sqrt{p_s-\hat{x}_2})\hat{x}_4]^2+[(-a_2\sqrt{x_3-p_r})x_4+(a_2\sqrt{\hat{x}_3-p_r})\hat{x}_4]^2}\le$$
$$\gamma\sqrt{(x_1-\hat{x}_1)^2+(x_2-\hat{x}_2)^2+(x_3-\hat{x}_3)^2+(x_4-\hat{x}_4)^2} \tag{7.11}$$

7.3 Design of Observer Gains

K and *q* in Eq. (7.6) are chosen by trial and error. Once *K* and *q* are chosen then, from Eq. (7.6), *p* is found. By using the inequality (7.5) the value of γ is calculated. This value of γ is then used to satisfy the inequality (7.11). It should be noticed that *K* and *q* must be chosen properly to determine a proper γ (Lipschitz constant).

$K=[450\ 400\ 100\ 0.52]^T$

$$q=\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

$$p = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 96.09\times10^{20} & 128.33\times10^{20} & 0 \\ 0 & 128.33\times10^{20} & 171.39\times10^{20} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$\lambda_{\min}(q) = 0.05$

$\lambda_{\max}(p) = 2.675\times10^{22}$

$\gamma = 9.346\times10^{-25}$

It was found that this value of γ is very small. Further trial and error for the selection of K and q will enable us to find a proper p, which will lead to a better γ to satisfy the inequality (7.11).

7.4 Fault Detection Methodology

Once the residuals are generated by the nonlinear observer, they are evaluated using the sequential test of Wald, also called the sequential probability ratio test, which is very useful in this research due to its characteristic feature that the number of observations required is not determined in advance. The primary objective of Wald's test is to observe the deviation between the output process and its estimates obtained by the observer called residuals or estimation errors. Both the system and observer output in this case are the velocity.

Figure (7.1) explains the fault detection scheme used in this thesis.

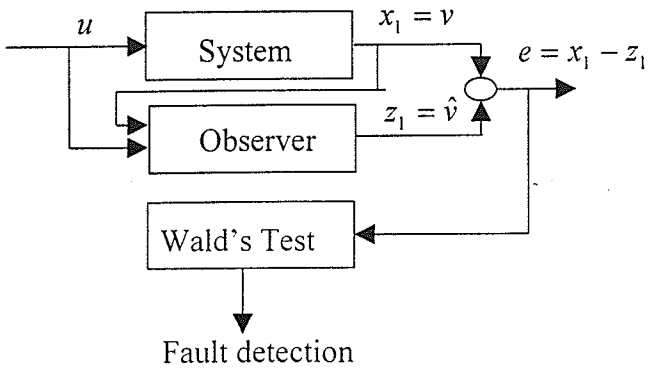


Fig. 7.1: Fault detection methodology.

In this context we face the situation where the analyzed values, i.e., the residuals, are distributed with an unknown mean, θ and a given standard deviation σ. From this point, the sequential test development takes into account the following. H_o is the hypothesis that the residual e under consideration is normally distributed with a mean θ_o. H_1 is the hypothesis that the residual e is normally distributed with a mean θ_1. In the sequential test, the hypothesis that $\theta = \theta_o$ is tested against the alternative that $\theta = \theta_1$. Let $e_1, e_2, ..., e_m$ be the successive observations on e. The probability density of the sample, $e_1, e_2, ...$ and e_m is given by

$$p_{om} = \frac{1}{(2\pi)^{\frac{m}{2}} \sigma^m} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(e_k - \theta_o)^2} \quad \text{if } \theta = \theta_o, \tag{7.12}$$

or

$$p_{1m} = \frac{1}{(2\pi)^{\frac{m}{2}} \sigma^m} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(e_k - \theta_1)^2} \quad \text{if } \theta = \theta_1. \tag{7.13}$$

The ratio $\frac{p_{1m}}{p_{om}}$ is computed at each stage of the test. Additional observations are taken as long as

$$B < \frac{p_{1m}}{p_{om}} = \frac{e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(e_k - \theta_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{a=1}^{m}(e_a - \theta_o)^2}} < A \tag{7.14}$$

Inspection is terminated with acceptance if

$$\frac{p_{1m}}{p_{om}} = \frac{e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(e_k - \theta_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(e_k - \theta_o)^2}} \leq B \tag{7.15}$$

Inspection is terminated with rejection if

$$\frac{p_{1m}}{p_{om}} = \frac{e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(x_k - \theta_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{m}(x_k - \theta_o)^2}} \geq A \tag{7.16}$$

Approximate values of A and B are given by $\frac{(1-\beta)}{\alpha}$ and $\frac{\beta}{(1-\alpha)}$, respectively, where α is the probability of false alarm (P_F) and β is the probability of missing alarm (P_M), i.e.,

$$P_F = P[\text{decide } H_1 | H_o \text{ is true}] \tag{7.17}$$

$$P_M = P[\text{decide } H_o | H_1 \text{ is true}] \tag{7.18}$$

By taking logarithms and simplifying inequalities (7.14), (7.15), and (7.16)

$$\ln\frac{\beta}{1-\alpha} < \frac{\theta_1-\theta_o}{\sigma^2}\sum_{k=1}^{m} e_k + \frac{m}{2\sigma^2}(\theta_0^2-\theta_1^2) < \ln\frac{1-\beta}{\alpha} \tag{7.19}$$

$$\frac{\theta_1-\theta_o}{\sigma^2}\sum_{k=1}^{m} e_k + \frac{m}{2\sigma^2}(\theta_0^2-\theta_1^2) \le \ln\frac{\beta}{1-\alpha} \tag{7.20}$$

and

$$\frac{\theta_1-\theta_o}{\sigma^2}\sum_{k=1}^{m} e_k + \frac{m}{2\sigma^2}(\theta_0^2-\theta_1^2) \ge \ln\frac{1-\beta}{\alpha} \tag{7.21}$$

Further simplification in carrying out the test procedure can be achieved by adding $(-\frac{m}{2\sigma^2})(\theta_o^2-\theta_1^2)$ to both sides of the inequalities (7.19), (7.20), and (7.21) and then dividing these inequalities by $\frac{(\theta_1-\theta_o)}{\sigma^2}$. These operations transform inequalities (7.19), (7.20), and (7.21) into

$$\frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\frac{\theta_o+\theta_1}{2} < \sum_{k=1}^{m} e_k < \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{1-\beta}{\alpha} + m\frac{\theta_o+\theta_1}{2} \tag{7.22}$$

$$\sum_{k=1}^{m} e_k \le \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\frac{\theta_o+\theta_1}{2} \tag{7.23}$$

$$\sum_{k=1}^{m} e_k \ge \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{1-\beta}{\alpha} + m\frac{\theta_o+\theta_1}{2} \tag{7.24}$$

By using the inequalities (7.22), (7.23), and (7.24), the inspection plan may be carried out as follows. For each m compute the acceptance number

$$a_m = \frac{\sigma^2}{\theta_1-\theta_o}\ln\frac{\beta}{1-\alpha} + m\frac{\theta_o+\theta_1}{2} \tag{7.25}$$

and the rejection number

$$r_m = \frac{\sigma^2}{\theta_1 - \theta_o} \ln \frac{1-\beta}{\alpha} + m \frac{\theta_o + \theta_1}{2} \tag{7.26}$$

These acceptance and rejection numbers are best computed before inspection starts. Inspection is continued as long as $a_m < \sum_{k=1}^{m} e_k < r_m$. Once $\sum_{k=1}^{m} e_k$ does not lie between a_m and r_m, inspection is terminated. There is no fault if $\sum_{k=1}^{m} e_k \le a_m$. Fault is detected if $\sum_{k=1}^{m} e_k \ge r_m$. The points ($m$, a_m) will lie on a straight line L_o and the points (m, r_m) will lie on a parallel line L_1. The common slope of the lines L_o and L_1 is given by

$$s = \frac{\theta_o + \theta_1}{2}$$

The intercept of L_o with y-axis is equal to

$$h_o = \frac{\sigma^2}{\theta_1 - \theta_o} \ln \frac{\beta}{1-\alpha}$$

and the intercept of L_1 with y-axis is equal to

$$h_1 = \frac{\sigma^2}{\theta_1 - \theta_o} \ln \frac{1-\beta}{\alpha}$$

8 Results

8.1 Simulation Studies for Observer

The observer was first tested in simulations. Simulation tests included sensitivity of the observer with respect to the changes in the model parameters.

8.1.1 Normal Operation

Observer results are shown for normal operation. The values of the parameters for system and observer are as follows:

$P_s = 100 \times 10^5$ Pa

$P_r = 0$

$f_d = 1600$N/m/s

$\beta = 5 \times 10^9$ Pa

$A_i = 0.002027\text{m}^2$

$A_o = 0.001520\text{m}^2$

$V_i = 0.015\text{m}^3$

$V_o = 0.015\text{m}^3$

$m = 20.0$kg

$\tau = 0.033$sec

$k_d = 0.032\text{m}^{3/2}/\text{kg}^{3/2}$

$w = 0.02$m

$K_1 = 450$

$K_2 = 400$

$K_3 = 100$

$K_4 = 0.52$

$k_{sp} = 0.00161$m/V

Figure 8.1 shows the control signal applied in simulations. Figure 8.2a shows both the observed and the actual velocities during extension and retraction of the actuator with different initial conditions. The values of initial conditions for the observer states are v = 0.3m/s, $P_i = 2 \times 10^5$Pa, $P_o = 2 \times 10^5$Pa and $X_{sp} = 0.001$m. In Fig. 8.2a the observed and the actual velocities converge asymptotically. In Figs. 8.2b and 8.2c both observed and actual pressure in and pressure out are shown with different initial conditions. Pressure is higher

during retraction than extension. This is due to the fact that both sides of the piston are not equal. During retraction, the working area of the piston is less, therefore, high pressure is required to generate enough force to cause retraction. During extension, the pressure is 3000kPa while during retraction, the pressure is 5000kPa.

8.1.2 Sensitivity Analysis

Observer sensitivity analysis is conducted with respect to the changes in the model parameters. Results for two parameters, bulk modulus and friction are presented as follows:

Bulk Modulus

The effect of any change in the system's bulk modulus on the observer performance is studied in simulations. When the bulk modulus is increased by a factor of 20, the observed velocity is able to converge asymptotically to the system's output velocity. But when bulk modulus is decreased by a factor of 20, the observed velocity cannot converge within 0.5 sec to the output velocity of the system as shown in Fig. 8.3a.

Figures 8.3b and 8.3c show the effects of increased and decreased bulk modulus of the system on observing the line pressures. The initial conditions for both the observer and the system are the same. It is seen that the line pressures are more sensitive to any change in the bulk modulus, particularly to the decreased bulk modulus. Figure 8.4 shows the effects of increased and decreased bulk modulus of the system on observer but with different initial conditions.

Viscous Friction

The effect of increased and decreased viscous friction on the performance of the observer is also studied. Figure 8.5a shows that when the value of viscous friction in the system is increased by a factor of 5, there is a slight difference between the observed and the actual velocities. Figure 8.5a shows that in the case of decreased friction, by a factor of 15, the

observer's and the system's output velocities converge asymptotically. Any change in friction has more impact on observed pressures (see Figs. 8.5b and 8.5c). Observed pressures are unable to converge asymptotically to the actual states. Figure 8.6 shows the observer performance when the initial conditions for both the observer and the system are different.

8.2 Simulation Studies for Fault Detection

Simulation studies are conducted for fault detection. Various faults, like cross-port and external leakages from the cylinder, incorrect pump pressure, faults due to changes in bulk modulus and sensor faults are studied. As discussed in Chapter 7, fault detection strategy consists of residual generation by using a nonlinear observer and evaluation of the residuals by using the sequential test of Wald, to detect the occurrence of fault. The following values are chosen for Wald's test:

$\beta = 0.03$

$\alpha = 0.03$

$\sigma = 0.1$

$\theta_o = 0.109$

$\theta_1 = 0.06$

In Fig. 8.7, two dotted lines show the thresholds computed by equations (7.25) and (7.26), while the straight line is the cumulative sum of the residuals which is the difference between observed and measured velocities. According to the criteria discussed in Section 7.4, if the cumulative sum of the errors is below the upper threshold, then there is no fault. Figure 8.7 shows that there is no fault. Any fault will generate residuals that are zero in this case. Figure 8.8a shows the occurrence of a fault. Fault has occurred due to incorrect supply pressure, which caused the deviation between the actual and observed velocity outputs of the system (The simulation model and the observer start with similar initial conditions). The difference is accumulated and plotted against the two thresholds (the dotted lines). Fault is detected when the cumulative sum of the residuals passes the upper threshold at the 170^{th} iteration. The time period between each iteration is

0.001seconds. It should be noted that this fault is due to a 30% increase in supply pump pressure. Figure 8.8b shows the result when increased supply pressure fault is removed after it is detected at 180^{th} iteration, The cumulative sum of the velocity errors enters in the region of no fault. Figure 8.9 shows the result pertaining to a 30% decrease in supply pump pressure fault. As can be seen, the fault is detected at the 190^{th} iteration.

Fault detection for cross-port and external leakages are shown in Fig. 8.10. The values of the leakage coefficients, C_i and C_o are 1×10^{-9}. How quickly the leakage fault can be detected depends upon the quantity of the leakage. More leakage will cause quicker detection of the fault.

In hydraulic applications, the bulk modulus of the liquid can change drastically due to the addition of air to the lines. The detection of a fault caused by a change in bulk modulus in the system is shown in Fig. 8.11. The bulk modulus of the system is reduced by a factor of 20, which affects the output of the system and causes the residuals to be generated. With reference to Fig. 8.11, fault is detected at the 25^{th} iteration due to the cumulative sum of the residuals crossing the upper threshold.

Sensor malfunction causes the control algorithms to be inefficient; it is important to detect a sensor fault promptly. Figure 8.12a shows the detection of the occurrence of a fault when the velocity sensor gain is increased by a factor of 1.4. Figure 8.12b shows the detection of the occurrence of a fault when the velocity sensor gain is decreased by a factor of 1.4. This was accomplished by multiplying the position sensor reading by 1.4.

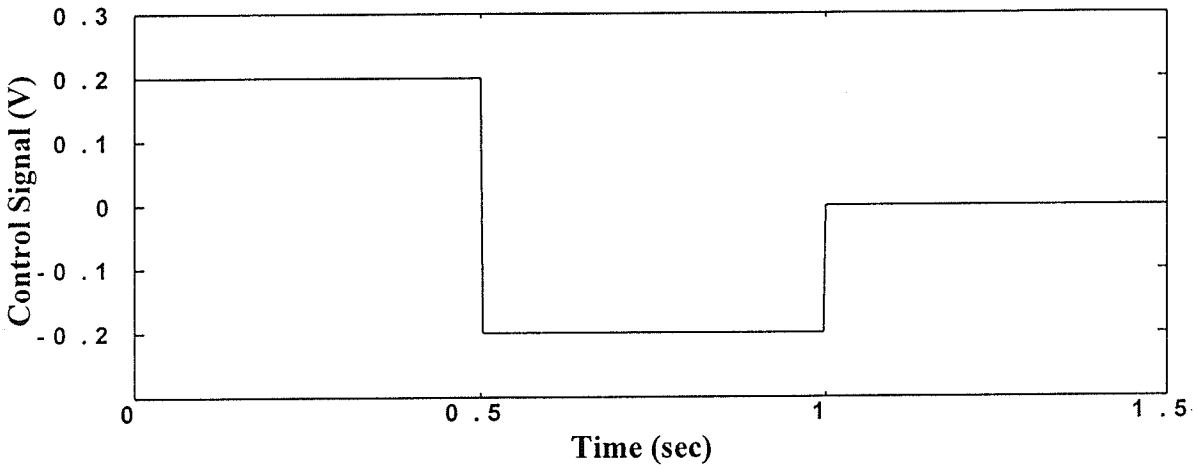


Fig. 8.1: Control signal.

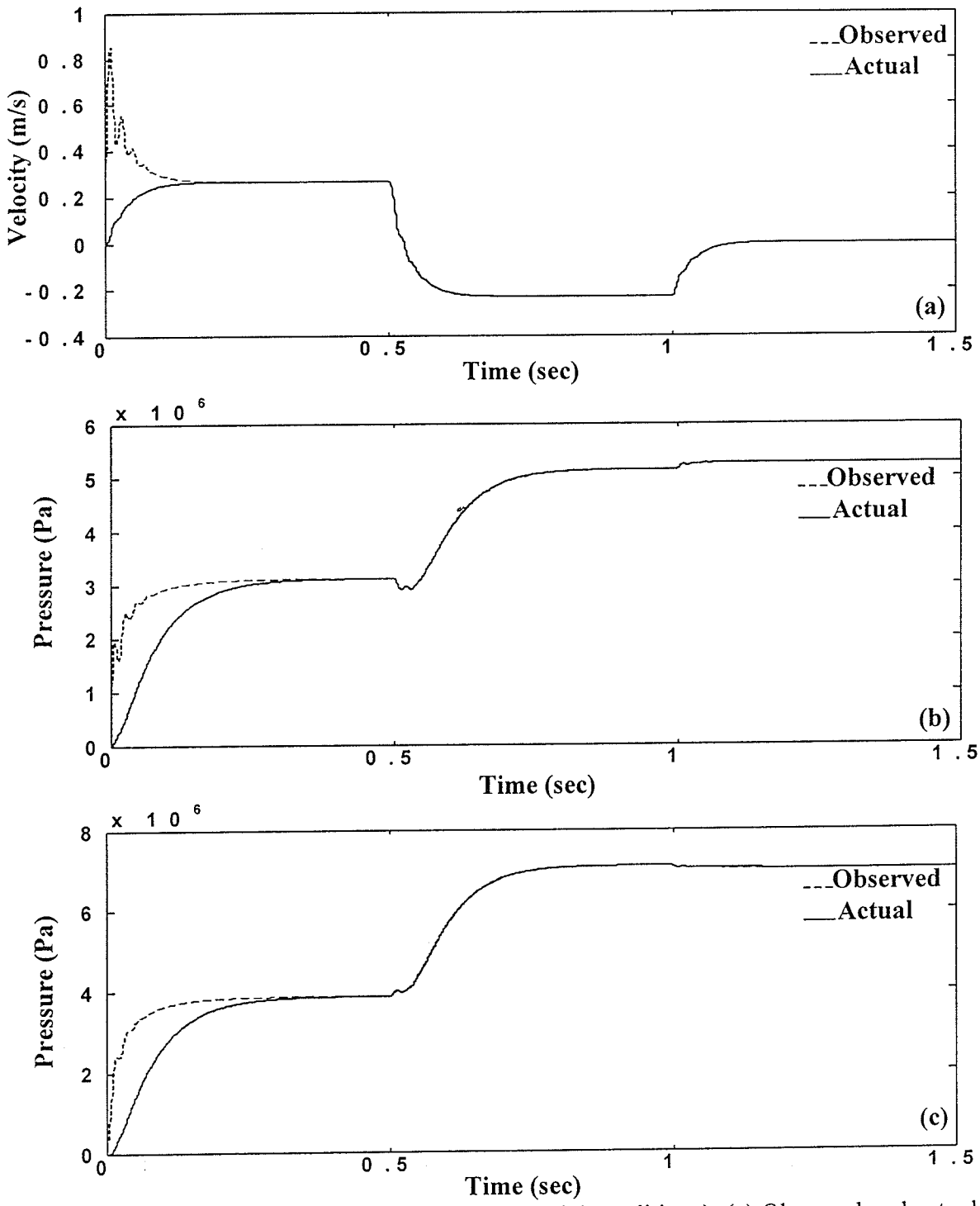


Fig. 8.2: Observer performance (different initial conditions): (a) Observed and actual velocities; (b) Observed and actual pressure in; (c) Observed and measured pressure out.

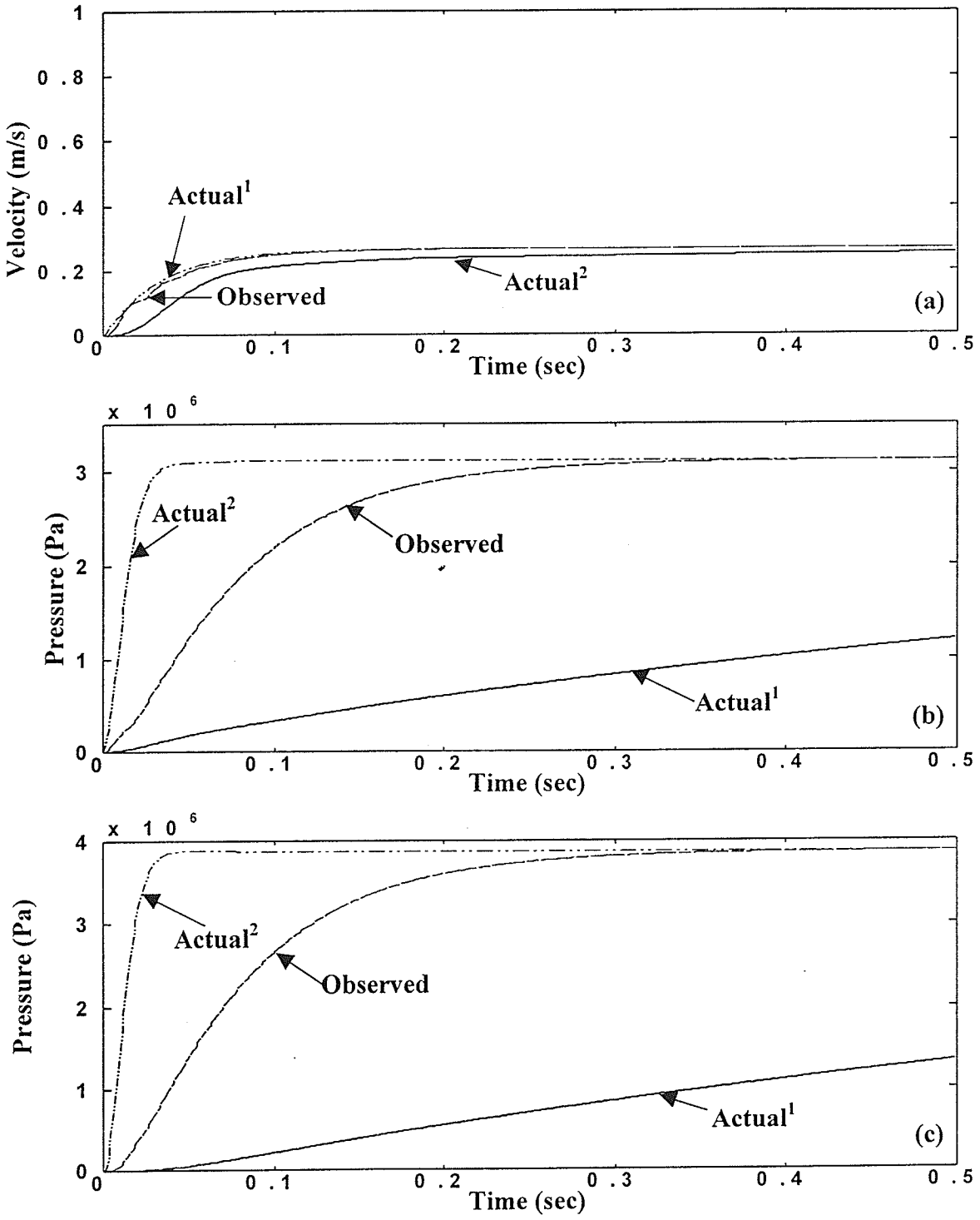


Fig. 8.3: Observer performance under changing bulk modulus (same initial conditions): (a)Observed and actual velocities; (b) Observed and actual pressure in; (c) Observed and actual pressure out.

Actual[1]: Bulk modulus decreased by factor 20

Actual[2]: Bulk modulus increased by factor 20

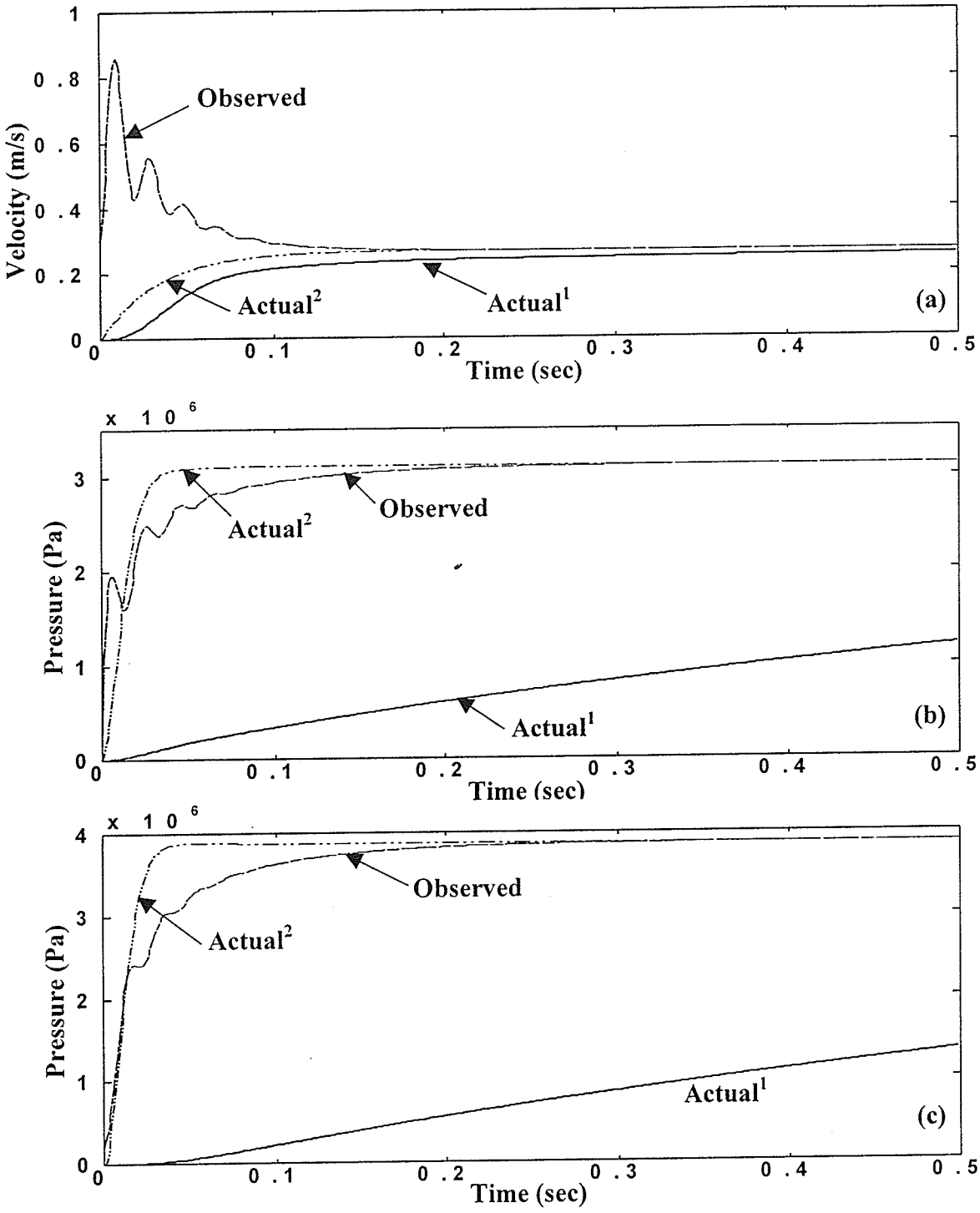


Fig. 8.4: Observer performance under changing bulk modulus (different initial conditions): (a) Observed and actual velocities; (b) Observed and actual pressure in; (c) Observed and actual pressure out.

Actual[1]: Bulk modulus decreased by factor 20

Actual[2]: Bulk modulus increased by factor 20

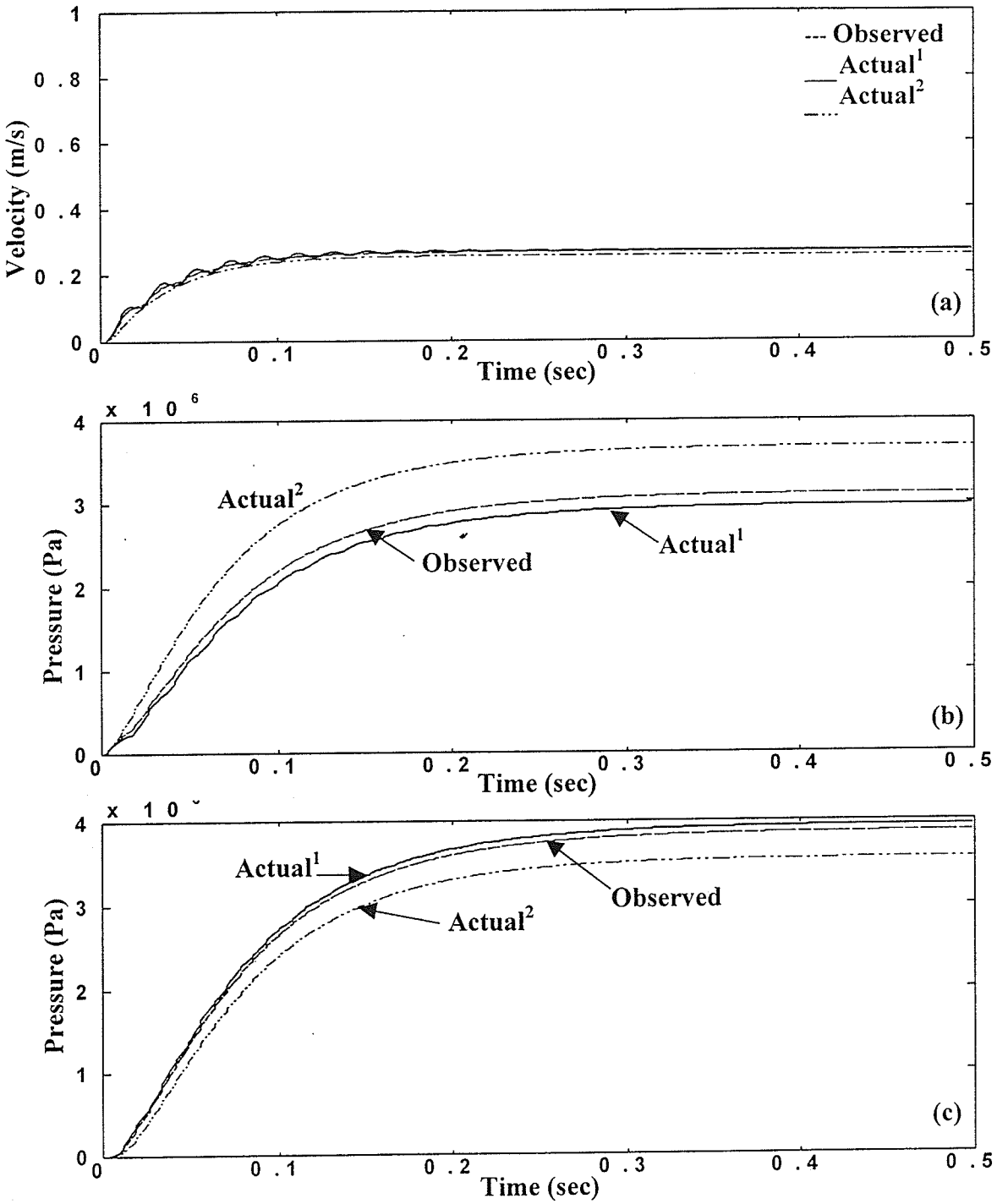


Fig. 8.5 Observer performance under changing friction (same initial conditions):
(a) Observed and actual velocities; (b) Observed and actual pressure in; (c) Observed and actual pressure out.

Actual[1]: Friction decreased 15 times

Actual[2]: Friction increased 5 times

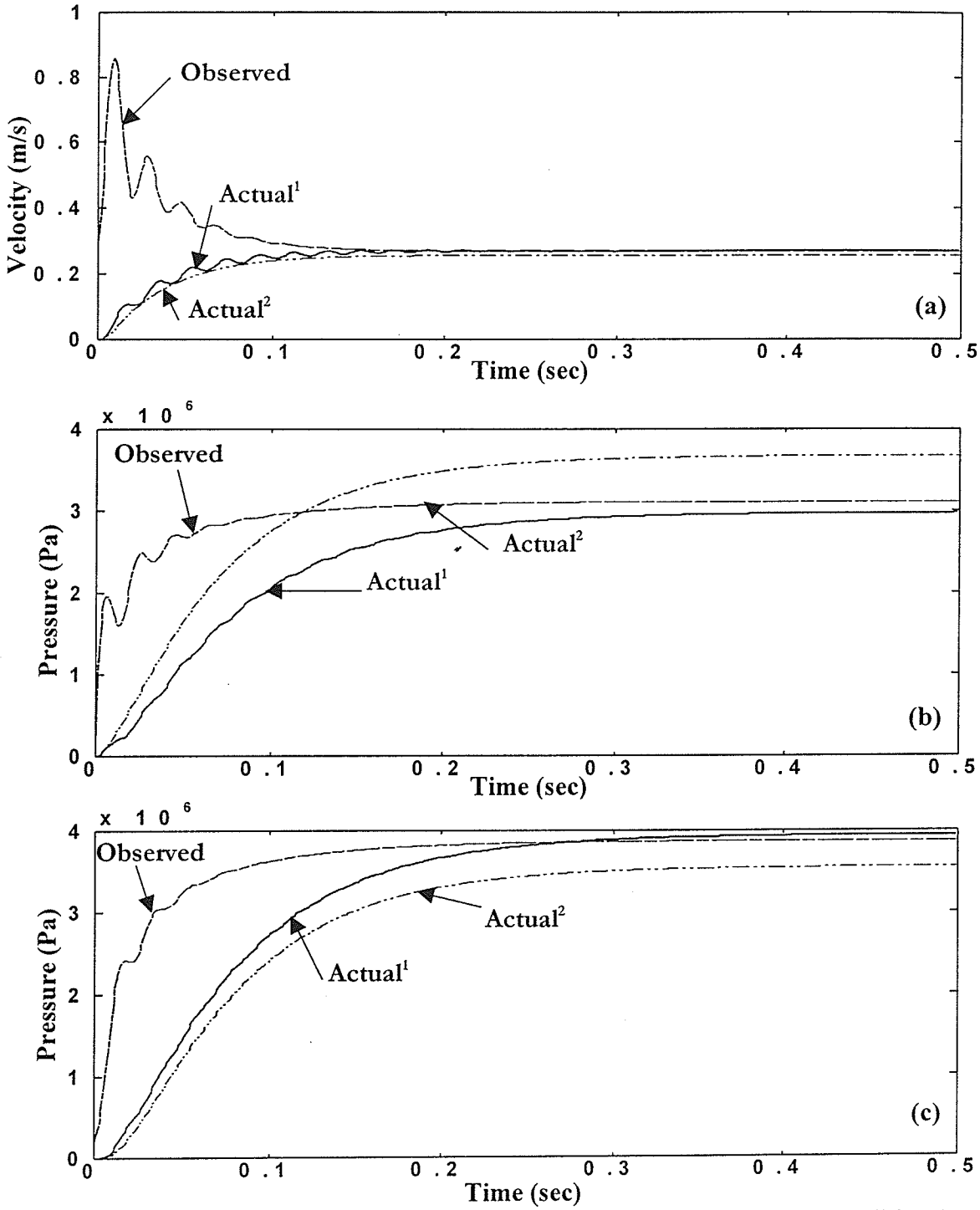


Fig. 8.6: Observer performance under changing friction (different initial conditions): (a) Observed and actual velocities; (b) Observed and actual pressure in; (c) Observed and actual pressure out.

Actual[1]: Friction decreased by factor 15

Actual[2]: Friction increased by factor 5

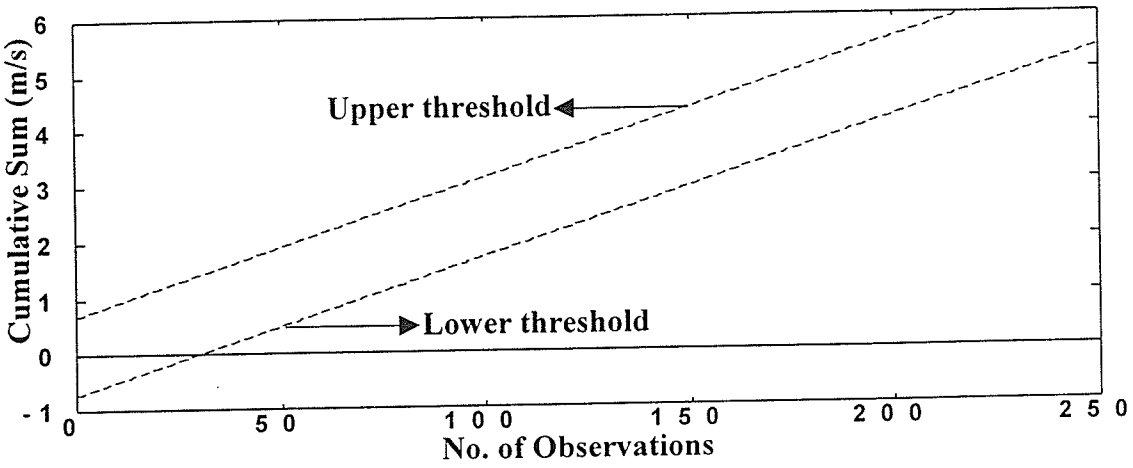


Fig. 8.7: Fault detection performance: Normal operation.

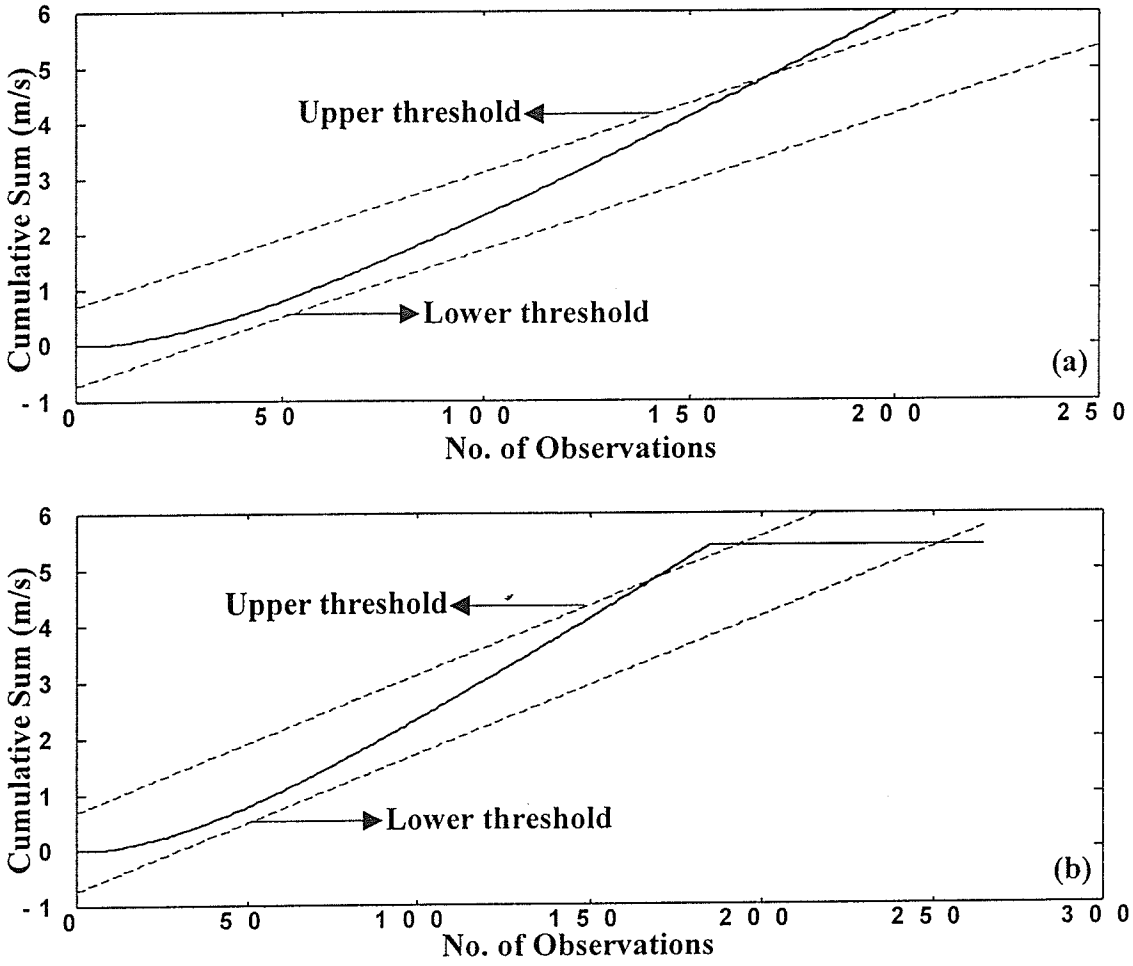


Fig. 8.8: Fault detection performance: (a) High pump pressure fault detection; (b) High pump pressure fault is removed after 180^{th} iteration.

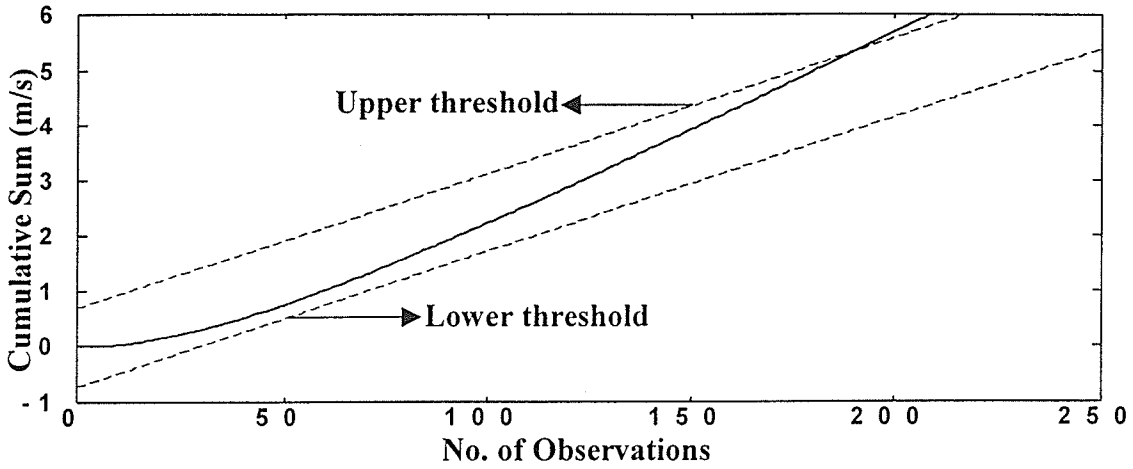


Fig. 8.9: Fault detection performance: Low pump pressure fault detection.

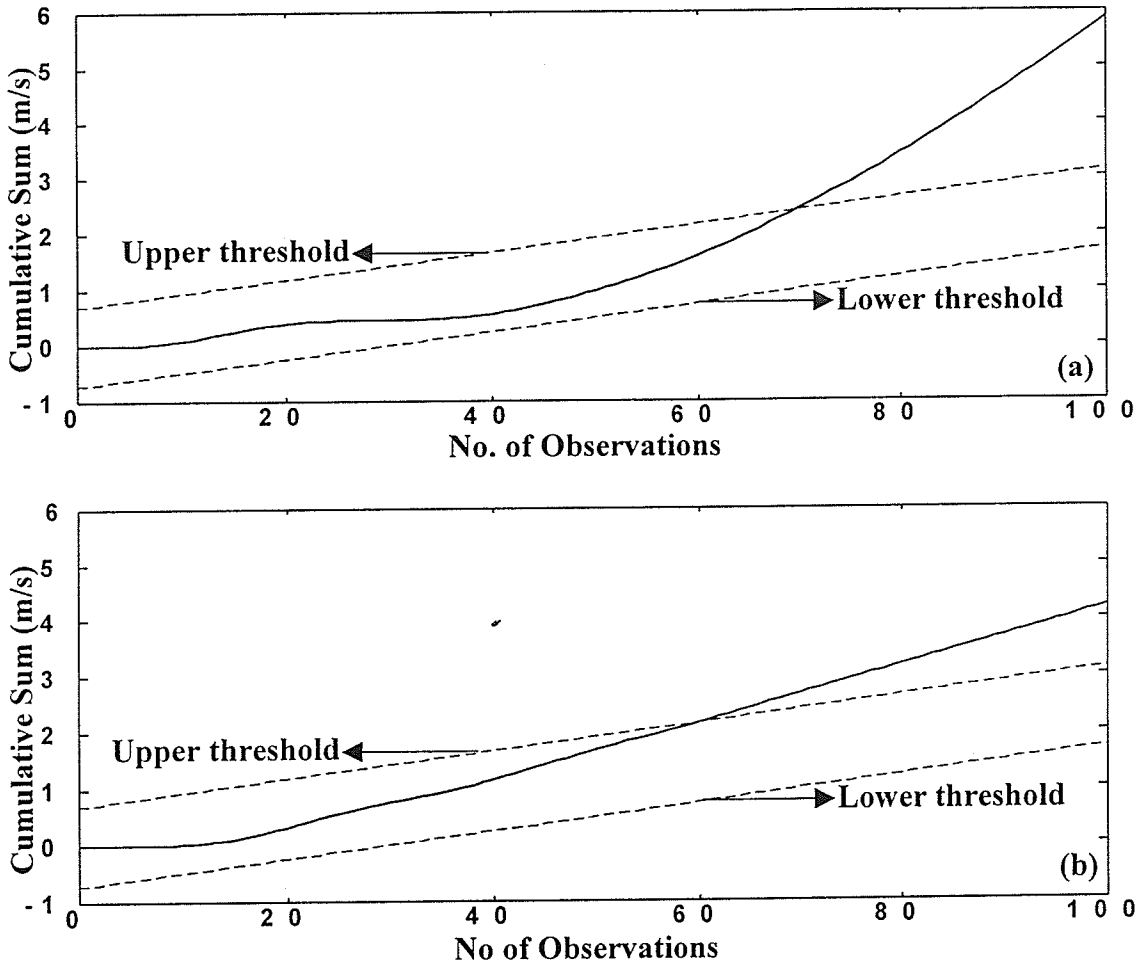


Fig. 8.10: Fault detection performance: (a) Cross-port leakage from the cylinder; (b) External leakage from the cylinder.

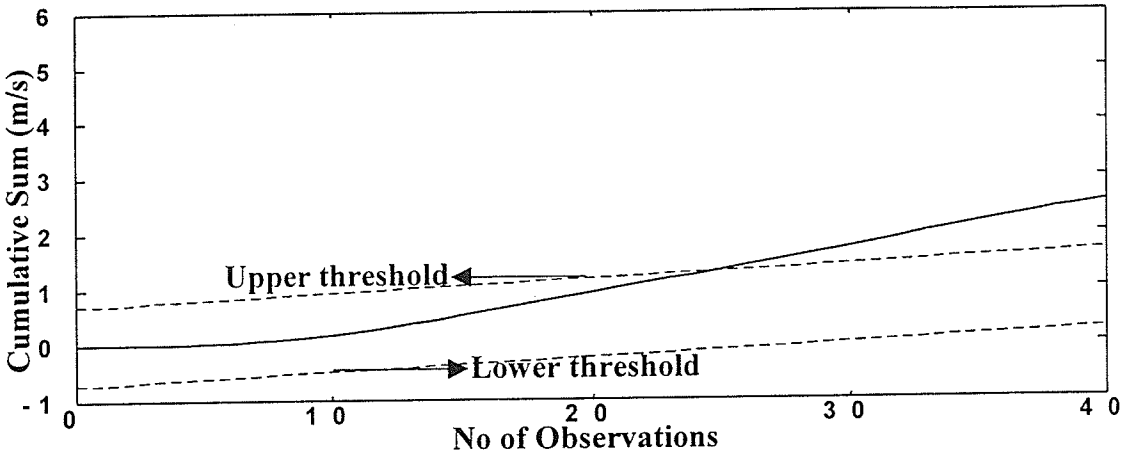


Fig. 8.11: Fault detection performance: Decreased bulk modulus fault detection.

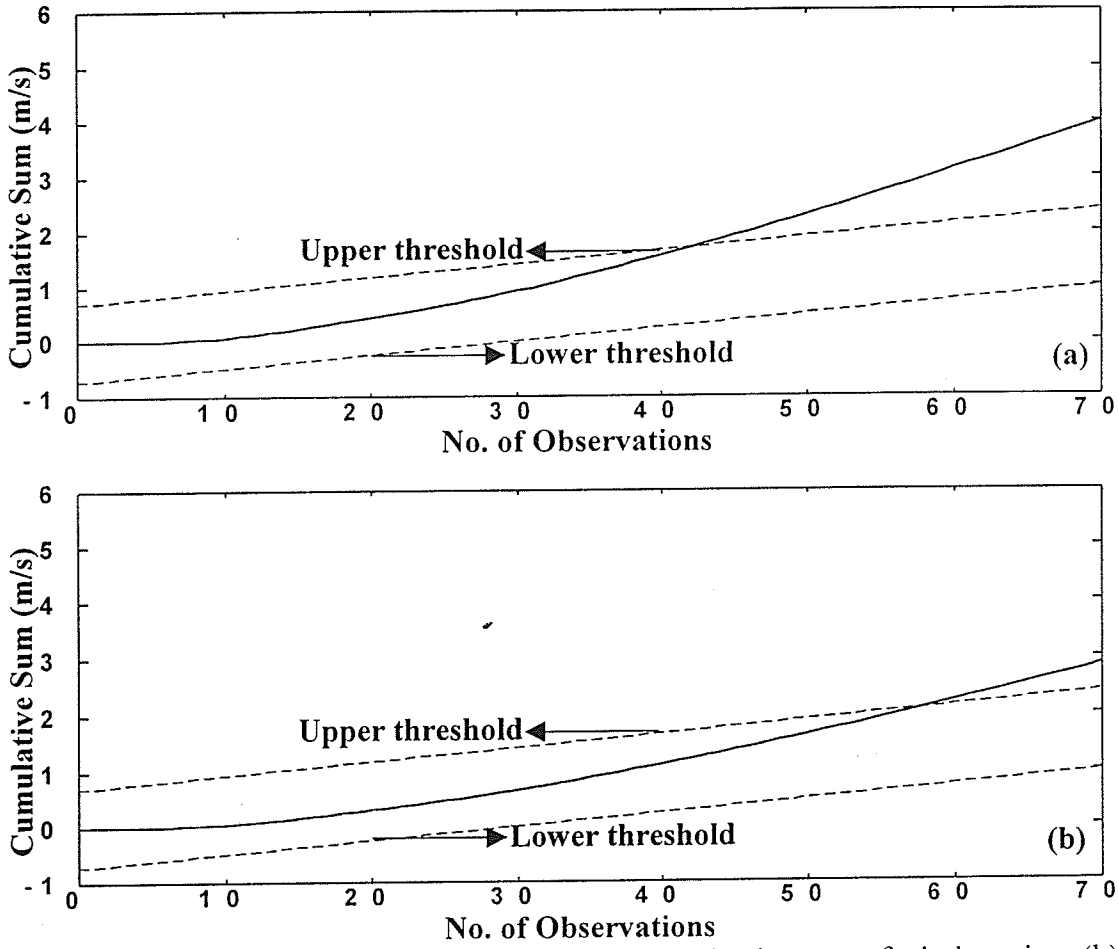


Fig. 8.12: Fault detection performance: (a) Increased gain sensor fault detection; (b) Decreased gain sensor fault detection.

8.3 Experimental Studies for Observer

Figure 8.13 shows the control signal, which is positive and causes positive spool valve displacement. Observer gains and other parameters are adjusted as follows:

$P_s = 7687\text{kPa}$ $\tau = 0.033\text{sec}$

$P_r = 0$ $k_d = 0.032\ \text{m}^{3/2}/\text{kg}^{3/2}$

$f_d = 1600\text{N/m/s}$ $w = 0.02\text{m}$

$\beta = 5\times10^9\ \text{Pa}$ $k_{sp} = 0.00161\text{m/V}$

$A_i = 0.00114\text{m}^2$ $K_1 = 410$

$A_o = 0.00063\text{m}^2$ $K_2 = 1.0\times10^4$

$V_i = 0.015\text{m}^3$ $K_2 = 1.0\times10^4$

$V_o = 0.015\text{m}^3$ $K_4 = 0.5$

$m = 20.0\text{kg}$

Figure 8.14a shows the two points velocity regression during the extension of the actuator. Velocity is quite noisy, which is not useful for further evaluation. For this reason 20 points regression is used, (see Fig. 8.14b) which produced an appropriate velocity profile. The time period between each iteration is 0.005 seconds.

Figure 8.15a shows both the output velocities from the observer and the real system during normal operation. There is a slight difference between the observed and measured velocities. This is due to the fact that there is always a difference between the model of the real system and the observer model. This difference, however should be minimized as much as possible by accurate modelling and by choosing the gains of the observer properly. Figure 8.15b shows the observed and measured input pressures. Observed pressure converges to the measured pressure. Figure 8.15c shows the observed and measured output pressures.

8.4 Experimental Studies for Fault Detection

Experimental studies were also conducted for fault detection. Two cases of faults were studied: (i) incorrect pump pressure fault, and (ii) sensor fault.

The values of Wald's test are as follows:

$\beta = 0.02$

$\alpha = 0.02$

$\sigma = 0.049$

$\theta_0 = 0.148$

$\theta_1 = 0.121$

Fig. 8.16 shows the result of fault detection when the system is under normal operation. Two dotted lines show the upper and lower thresholds. The cumulative sum of the residuals, that is, the difference between the measured and observed velocities, is the solid line which lies between the two thresholds. Since the cumulative sum of the residuals is not crossing the upper threshold, no fault is detected and the system is considered to be working in a normal operational mode. By comparing Figs. 8.7 and 8.16, it is seen that the cumulative sum of the residuals is zero in the simulation, whereas, it is not zero during the experiment. This is believed to be due to the difference between the values of the parameters used in the observer and those belonging to the actual system. Therefore, residual generation occurs even during the normal operation of the experimental system. Nevertheless, as long as the cumulative sum of the residuals is below the upper threshold, there is no fault and the system is considered to be working normal.

When the pump pressure drops due to a fault from a normal operating pump pressure of 7687kPa ($\approx$ 1100 psi) to 6357kPa ($\approx$ 900 psi) (see Fig. 8.17a), the cumulative sum of the residuals crosses the upper threshold at the 700th iteration, confirming the occurrence of the fault. Figure 8.17b shows the performance of the fault detection technique when the pump pressure is dropped from 7687kPa ($\approx$ 1100 psi) to 5667kPa ($\approx$ 800 psi). The fault is detected earlier as compared to the first case.

Figures 8.17c, 8.17d and 8.17e show the detection of the faults when the supply pressure drops to 4964kPa ($\approx$ 700 psi), 4309kPa ($\approx$ 600 psi) and 3585kPa ($\approx$ 500 psi), respectively. It should be noted that the greater the pump pressure fault, the sooner the detection of the fault. Figure 8.18 shows the position sensor fault detection (sensor gain is decreased by a factor of 1.3). The fault is detected at the 100th iteration.

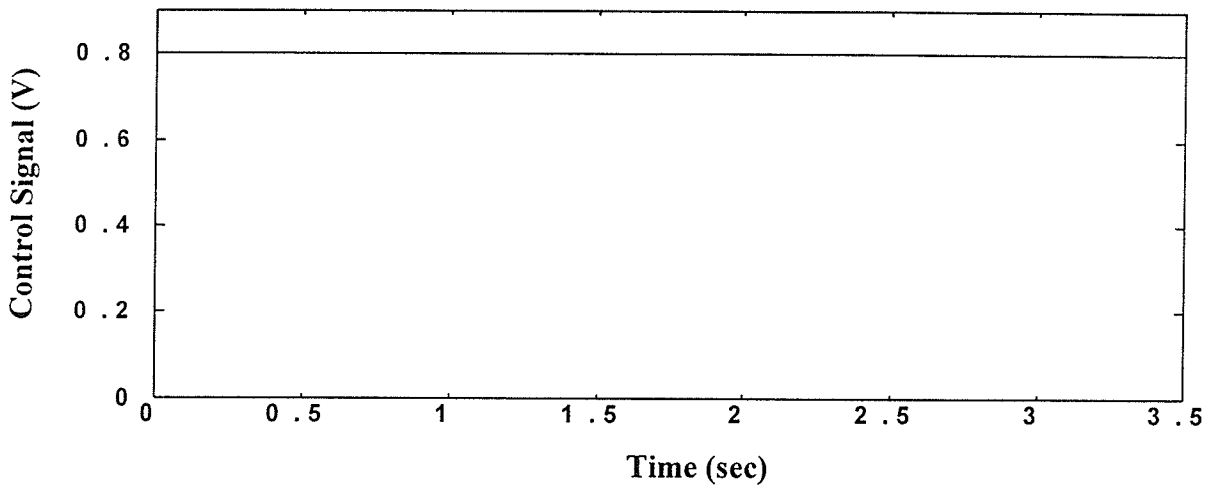


Fig. 8.13: Control signal.

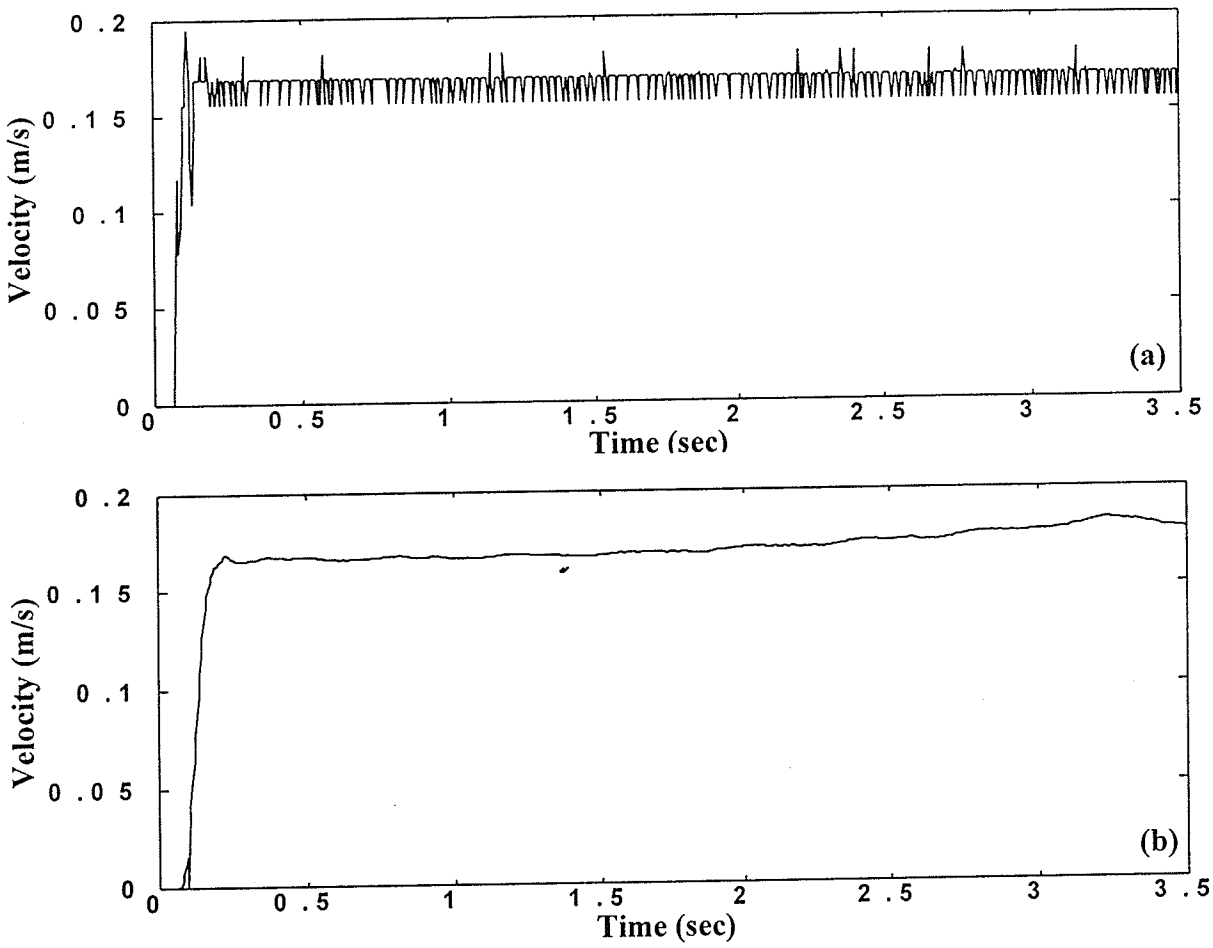


Fig. 8.14: Observer performance: (a) Velocity, two points regression; (b) Velocity, 20 points regression.

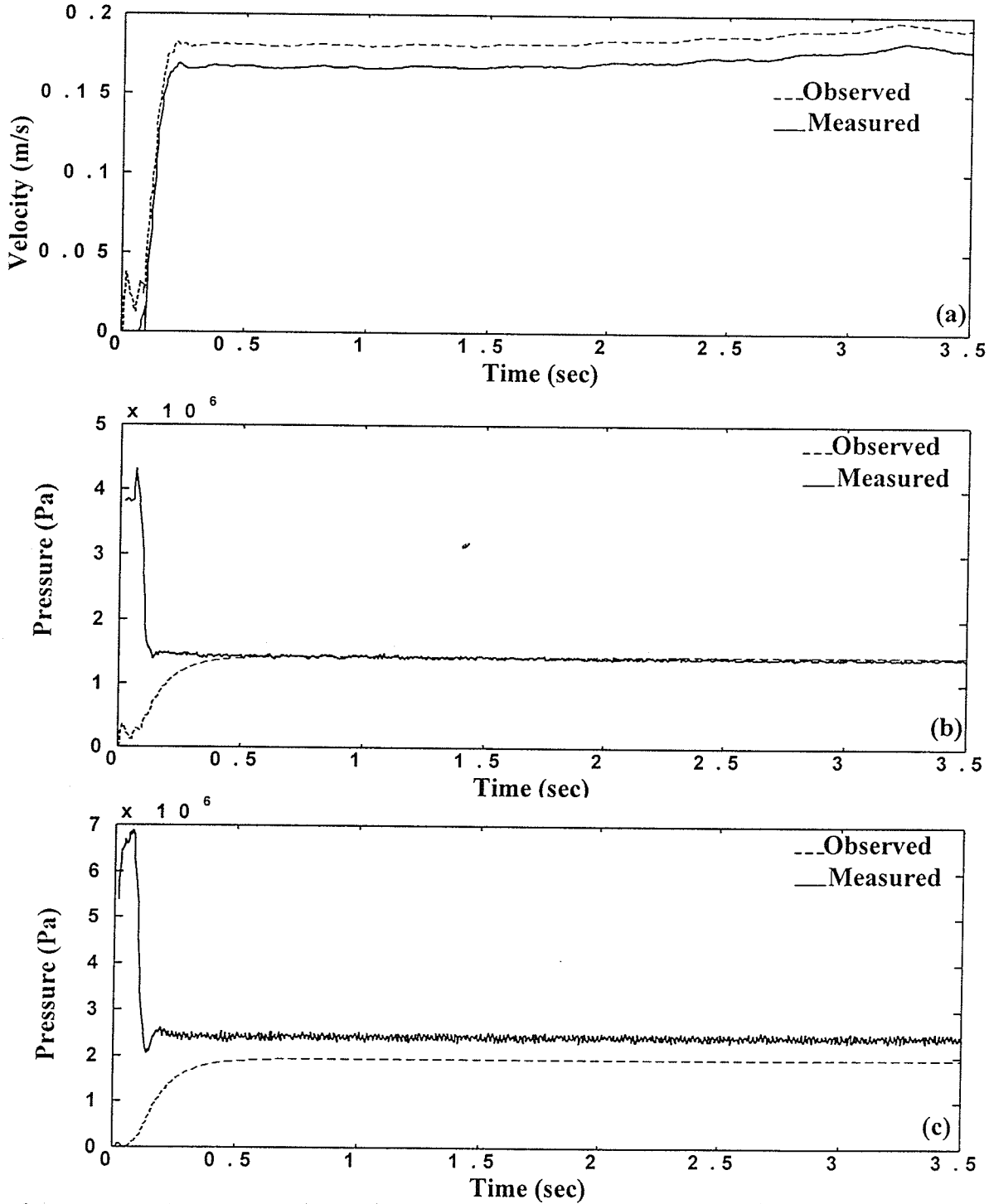


Fig. 8.15: Observer Performance: (a) Observed and measured velocities; (b) Observed and measured pressure in; (c) Observed and measured pressure out.

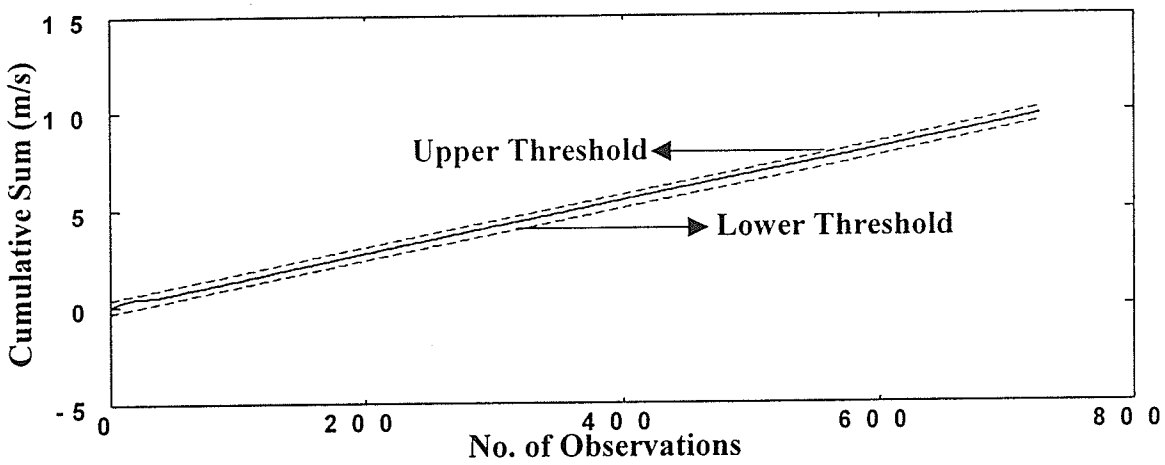
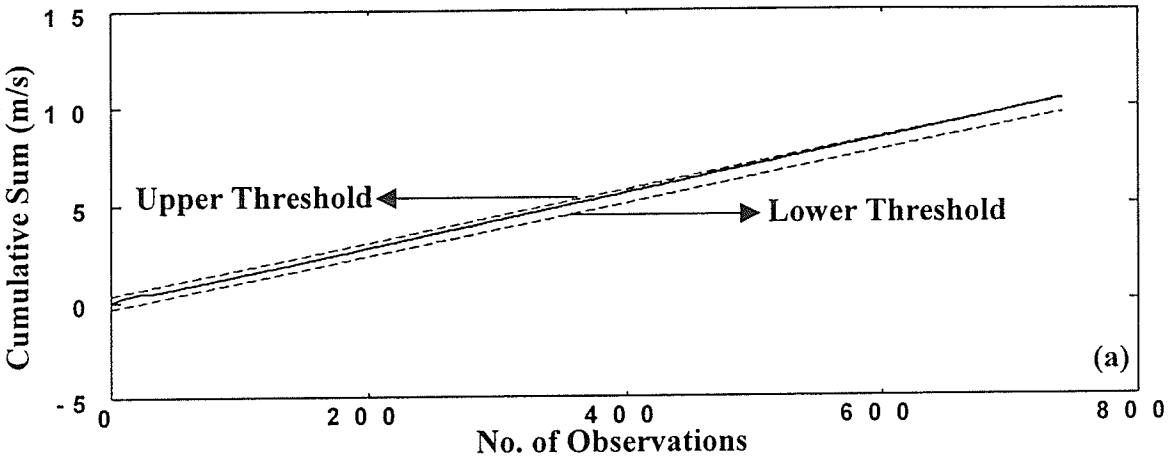
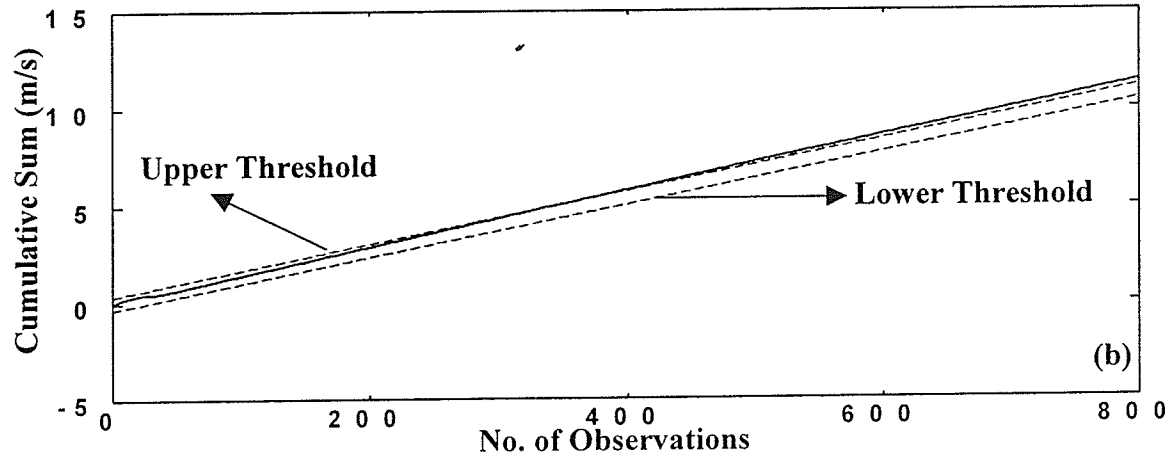
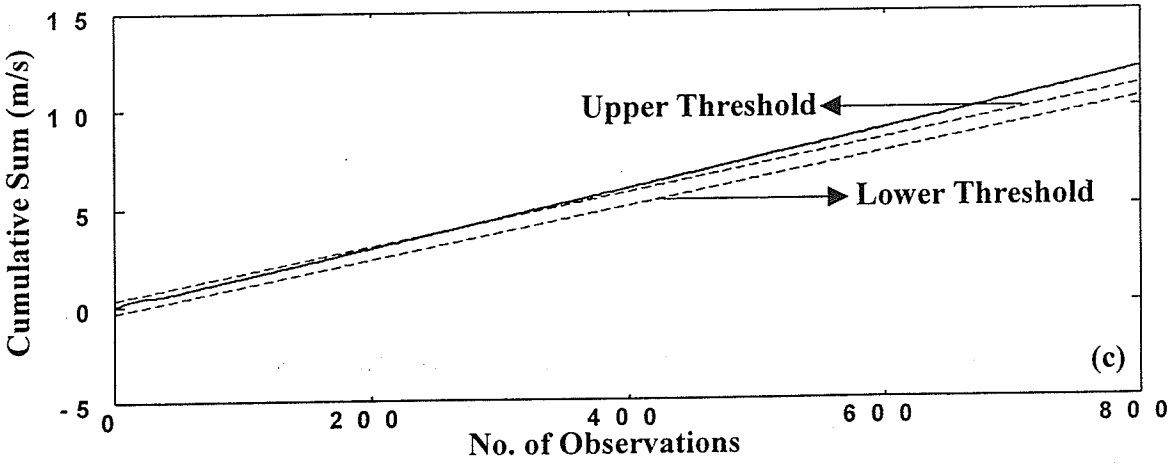


Fig. 8.16: Fault detection performance: Normal operation.


Cumulative Sum (m/s)
15
10
5
0
-5
Upper Threshold
Lower Threshold
(a)
0
200
400
600
800
No. of Observations


Cumulative Sum (m/s)
15
10
5
0
-5
Upper Threshold
Lower Threshold
(b)
0
200
400
600
800
No. of Observations


Cumulative Sum (m/s)
15
10
5
0
-5
Upper Threshold
Lower Threshold
(c)
0
200
400
600
800
No. of Observations

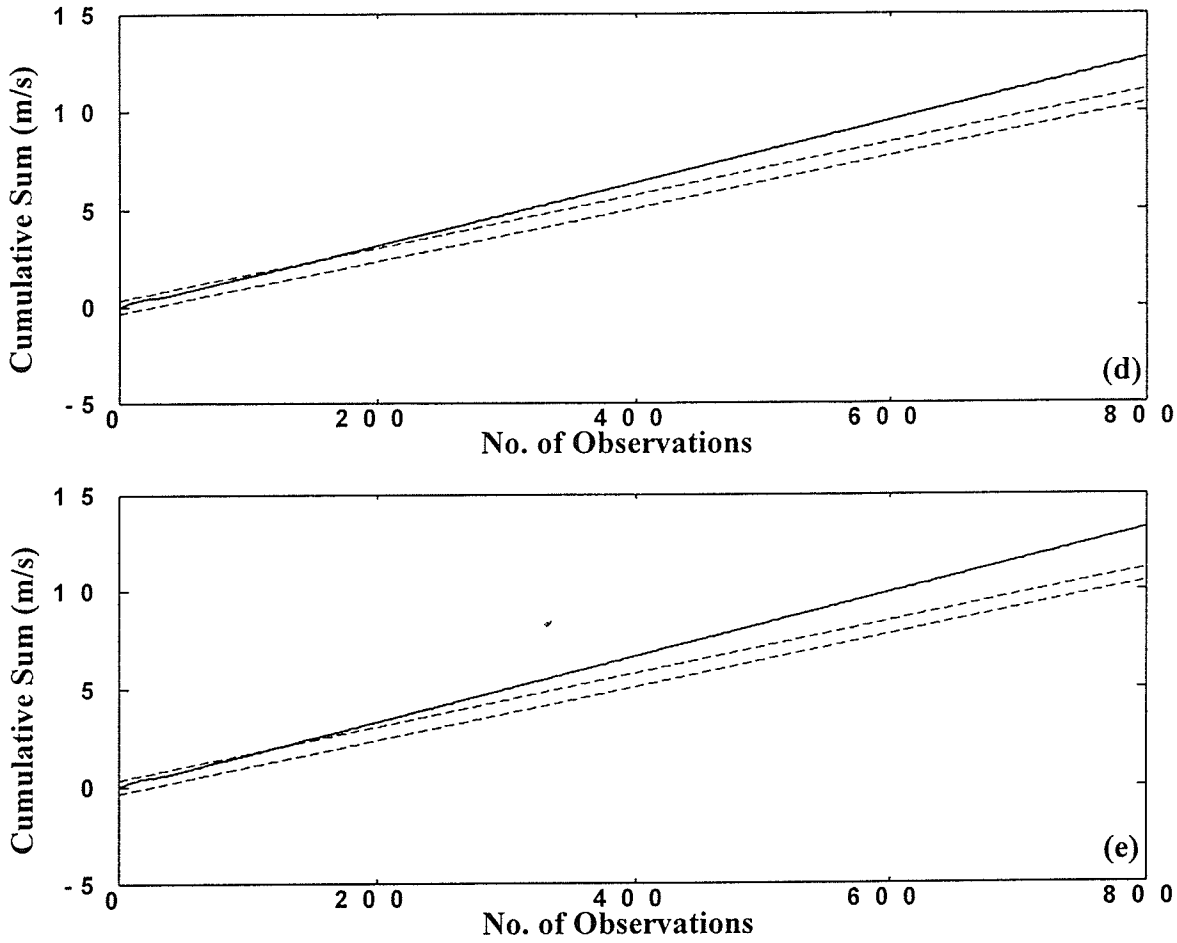


Fig. 8.17: Fault detection performance: (a) Fault detected at lower pressure of 6357kPa; (b) Fault detected at lower pressure of 5667kPa; (c) Fault detected at lower pressure of 4964kPa; (d) Fault detected at lower pressure of 4309kPa; (e) Fault detected at lower pressure of 3585kPa.

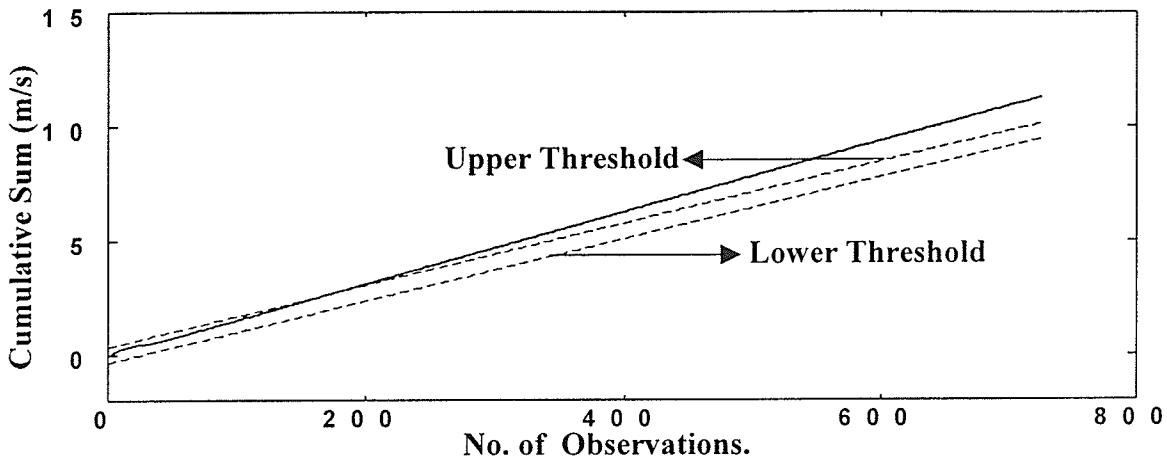


Fig. 8.18: Position sensor fault detection.

9 Conclusions

The achievements of this research are two-fold: (i) employing a nonlinear observer for the class of electro-hydraulic servo-positioning systems, and (ii) incorporating the sequential test of Wald to detect the occurrence of a fault. A nonlinear observer is proposed based on the Lipschitz class of nonlinear equations. The observer is driven by the input signal and the output state of the system, which is the velocity. The observer also estimates the other states such as the line pressures. The residuals are taken from the difference of the measured and the observed velocities and are further evaluated by the sequential test of Wald to detect the occurrence of a fault.

Since the nonlinearity in the system under investigation comes from the valve flow equations, it was convenient to develop a nonlinear observer. Simulation studies were conducted to observe the sensitivity of the observer with respect to viscous friction and bulk modulus changes in the system. It was found that the observer is more sensitive to changes in the bulk modulus than to changes in other parameters of the system. All results show good performance of the nonlinear observer.

Simulation studies were also conducted using Wald's sequential test for fault detection. Simulation results show the effectiveness of the fault detection strategy. Since the detection method utilized residuals, the residuals were generated successfully for all types of common hydraulic system faults, i.e., incorrect pump pressure, cross-port leakage, external leakage from the cylinder, changes in the bulk modulus and sensor faults. It was shown that residual generation is most difficult in the case of increased bulk modulus due to the small error between actual and the observed outputs.

Experimental verifications were conducted for both observer and fault detection. It was shown that observed states were adequately following the actual states. This proved the good performance of the nonlinear observer. Experimental verifications were also conducted for two types of faults: (i) incorrect pump pressure and (ii) sensor faults. It was

not possible to conduct the experiment for further faults due to the limitations of the experimental test rig.

10 References

Adjallah, K., Maquin, D. and RAGOT, J., 1994, "Nonlinear Observer-Based Fault Detection", *IEEE Conference on Control Applications*, Vol. 2, pp. 1115-1120.

Beard, R.V., 1971, "Failure Accommodation in Linear Systems Through Self-Reorganization", *Man-Vehicle Lab. Mass. Inst. Technology, Cambridge*, MA, Rep. MVT-71-1

Crowther, W. J., Edge, K. A., Burrows C. R., Atkinson, R. M., and Woollons, D. J., 1998, "Fault Diagnosis of a Hydraulic Actuator Circuit Using Neural Networks and Output Vector Space Classification Approach", *I. Mech. E., Proc. Instn. Mech. Engrs.* Vol. 212 part 1.

Gertler, J., 1998, "*Fault Detection and Diagnosis in Engineering Systems*", Marcell Dekker, Inc. New York.

Hengy, D. and Frank, P.M., 1986, "Component Failure Detection via Nonlinear State Observers", *IFAC Workshop Kyoto,* Japan.

Himmelblau, D.M., 1978, "*Fault Detection and Diagnosis in Chemical and Petrochemical Processes",* Elsevier Scientific Publishing Company. Amsterdam.

Hou, M. and Muller, P., 1992, "Design of Observers for Linear Systems With Unknown Inputs", *IEEE Transcations on Automatic Control*, Vol. 37, No. 6, pp. 871-875.

Hou, M. and Muller, P., 1991, "Design of Robust Observers for Fault Isolation", *Proceedings of the IFAC/IMACS Symposyium on Fault Detection, Supervision and Safety for Technical Processes- SAFEPROCESS '91, Baden-Baden, Germany*, pp. 265-270.

Isermann, R., 1984, "Process Fault Detection Based On Modelling And Estimation Methods, A Survey", *Automatica*, Vol.20, No.4, pp 387-404.

Jones, H.L., 1973, "Failure Detection in Linear Systems", *Ph.D. dissertation, Dep. Aeronautics and Astronautics, Mass. Inst. Technology.*

Luenberger, D. G., 1971, "An Introduction to Observers", *IEEE Transaction on Automatic Control*, Vol. AC-16, No.6, pp. 596-602.

Massoumnia, M.A., 1986, "A Geometric Approach to The Synthesis of Failure Detection Filters", *IEEE Transactions on Automatic Control*, Vol. AC-31, No. 9, pp. 839-846.

Merrit, H.E., 1967, *Hydraulic Control Systems.* Wiley, New York.

Patton, R., Frank, P.M., and Clark, R., 1989, *Fault Diagnosis in Dynamic Systems, Theory and Applications*. Prentice Hall International Ltd. (UK).

Patton, R.J. and Chen, J., 1991, "Robust Fault Detection Using Eigenstructure Assignment: A Tutorial Consideration and Some New Results", *Proceedings of the 30th Conference of Decision and Control,* Brighton, England, pp. 2242-2247.

Rajamani, R., 1998, "Observers for Lipschitz Nonlinear Systems", *IEEE Transactions On Automatic Control,* Vol. 43, No. 3, pp. 397-401.

Rajamani, R., Hedrick, J. K., 1995, "Adaptive Observers for Active Automotive Suspensions: Theory and Experiment", *IEEE Transactions on Control Systems Technology,* Vol. 2, No. 1, pp 86-93.

Seliger, R. and Frank, P.M., 1991, "Rŏbust Component Fault Detection And Isolation In Nonlinear Dynamic Systems Using Nonlinear Unknown Input Observers", *Proceedings of the IFAC/IMACS Symposyium on Fault Detection, Supervision and Safety for Technical Processes – SAFEPROCESS '91, Baden-Baden, Germany*, pp. 313-318.

Thau, F. E., 1973, "Observing the State of Nonlinear Dynamic Systems", *International Journal of Control,* Vol. 17, No. 3, pp. 471-479.

Viswanadham and Srichander, R., 1987, "Fault Detection Using Unknown Input Observer", *Control Theory and Advanced Technology*, Vol. 3, No. 2, pp. 91-101.

Wald, A., 1947, *Sequential Analysis*. Dover Publications, Inc. New York.

Watton, J., Lucca-Negro, O., and Stewart, J. C., 1994, "An On-line Approach to Fault Diagnosis of Fluid Power Cylinder Drive Systems", *Proc. Institution of Mechanical Engineers, Part I, Journal of Systems and Control Engineering,* Vol. 208, No. 14, pp. 249-262.

White, J.E. and Speyer, J.L., 1987, "Detection Filter Design: Spectral Theory and Algorithms", *IEEE Transactions on Automatic Control*, Vol. AC-32, No. 7, pp. 593-603.

Yu, D., 2000, "Diagnosing Simulated Faults for an Industrial Furnace Based on Bilinear Model", *IEEE Transactions on Control Systems Technology,* Vol. 8, No. 2, pp. 435-442.

Yu, D., Shields, D.N., and Mahatani, J.L., 1994, " A Nonlinear Fault Detection Method for a Hydraulic System", *IEE Control Conference,* Publication No. 389, pp. 1318-1322.

Yuksel, Y. O., and Bongiorno, J. J., 1971, "Observers for Linear Multivariable Systems with Applications", *IEEE Transactions on Automatic Control*, Vol. 16, No. 6, pp. 603-612.