CLOSED-LOOP IDENTIFICATION

USING QUANTIZED DATA

A thesis submitted to the University of London

for the degree of

Doctor of Philosophy

by

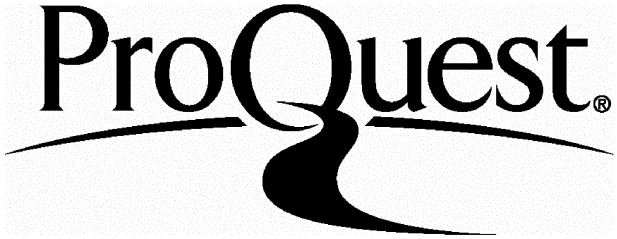
Meihong Wang

University College London

Department of Electronic and Electrical Engineering

Torrington Place London WC1E 7JE

November 2002



To Zhongqin, Rudai and Rumeng

ACKNOWLEDGEMENTS

It is thanks to Nina Thornhill that these three years as a PhD student in University College London and Imperial College London have been most enjoyable. Thank you for consistent encouragement, support and help, and for being a wonderful thesis supervisor.

I would like to acknowledge The Centre for Process Systems Engineering in Imperial College of Science, Technology and Medicine for sponsoring this research and funding me for three years. The Centre is really a nice academic environment for PhD research with so many research active staff and students such as Vassilis Kosmidis, Vassilis Sakizlis and Marcus Vinicius and Thomas Apelt. Thank you for the helpful discussions and the social talks.

Thanks also go to Sirish Shah, Biao Huang and Bhushan Gopaluni of the Computer Process Control group in The University of Alberta in Canada for the permission to use and help on using the pilot plant for the experimental work during the days when I was there.

Last, but not least, I would like to thank my friends Yong Yiu and Jie Zhang, Keming Ma and Haitao Huang for their kind care and support throughout these years.

ABSTRACT

A model of a system is important for applications such as simulation, prediction and control. Closed-loop identification (CLID) is a means of identifying a process model while the process is still under feedback control. The motivation of this project is to find a way to do closed-loop identification while causing minimum disruption to the controlled process.

There are two main categories of closed loop identification. One is closed-loop identification with external excitation (Ljung 1987, *System Identification: theory for the user*, Englewood Cliffs, NJ: Prentice-Hall). Another is relay identification (Åström and Hägglund 1984a, Automatic Tuning of Simple Regulators with Specifications on Phase and Amplitude Margins, *Automatica*, Vol.20, No.5, pp645-651). The first achievement of this thesis is the establishment of a connection between previously unrelated facts by comparing the two main categories of closed loop identification methods. Their advantages and disadvantages were highlighted through case studies.

The second, and the main achievement of this thesis is to propose a new closed-loop identification scheme for a single-input-single-output (SISO) control loop. It is based on a quantizer inserted into the feedback path. The novel contribution of this thesis is to bring the closed-loop identification with external excitation method and the relay identification method into a unified framework for the first time. It gives recommendations about the appropriate method to use for a given quantizer interval. When the quantization interval is small, the quantization error is persistently exciting, equivalent to an external excitation. The two-stage (step) method can be applied. When the quantization interval is large, the relay method can be applied instead. Nonlinearity caused by the quantizer is analyzed, which indicates that nonlinearity increases with the quantization interval. Simulations and experiments showed that the proposed closed-loop identification scheme based on quantization is successful.

The third achievement of this thesis is the implementation, testing and extension of a quantized regression (QR) algorithm that retrieves the underlying information from quantized signals such as those from the analogue to digital converter of a plant instrument. The algorithm is a combination of the 'Gaussian Fit' scheme with expectation-maximization (EM) algorithm. The new QR algorithm can optimally estimate the model parameters and recover the underlying signal at the same time for an arbitrary number of quantizer levels.

Table of Contents

2 LITERATURE REVIEW

List of Figures

List of Tables

Nomenclature

The nomenclature follows Soderstrom & Stoica (1989) and/or Ljung (1987)

$\hat{\theta}$: an estimate

θ_0: the true value

$E\hat{\theta}$: E is the expectation operation

N: the number of data points used in the estimate

Z^N: the data set

$\mathcal{M}$: Model Structure

$\hat{G}$: an estimate of the true process G_0

G_0: the true process

C: the controller

d: the relay amplitude in the context of relay-based identification

$d(t)$: the external excitation in the context of closed loop identification with external excitation.

a: the amplitude of the oscillation in the process output (limit cycle) in the context of relay-based identification

$a(t)$: the noise in the context of closed loop identification with external excitation

K: the process steady-state gain

L: the dead time

T: the time constant

s: Laplace variable

qi: quantization interval

ql: quantization level

h: quantizer output amplitude

ε: hysteresis width

u: process input

y: the process output

y_{sp}: the setpoint

CHAPTER 1

INTRODUCTION

This chapter begins by presenting an overview, followed by the important definitions and concepts in system identification and closed-loop system identification. The objective and the motivation of this thesis and the key issues to be addressed are described. The last sections give the outline of the approaches used throughout this thesis and the organization of the thesis.

1.1 Overview

1.1.1 An example

The closed loop control system is common in everyday life, for example a hot water tank in a domestic house. In this example, the hot water tank is the controlled system because it is desired to maintain the hot water at a desired temperature by means of a heating element. The tank and the heater comprise a closed loop. There are generally two ways to find the model of a system. One way is through application of the laws of physics. The other way is by system identification. For the hot tank example, the first possible approach is to calculate the model after knowing the size of the tank, the heat transfer rates and so on. Another is to infer the model from measurements of the heat input and temperature over a period of time. Closed-loop identification uses the latter method. Closed-loop identification is a challenge because the aim of the control system is to keep the temperature constant, yet identification requires the temperature to change in order to give some information about the tank.

1.1.2 Motivation

This thesis aims to find novel ways to meet the challenges involved in closed-loop identification. For the hot tank example, on the one hand, the system must be disturbed, for instance, by switching the heater on and off. On the other hand, the disturbance must be minimized in order to maintain the temperature at the correct value. Therefore the motivation of the work is to find a means of supplying sufficient information with a minimum of disturbance.

Quantization caused by analogue to digital (A/D) conversion and/or digital to analogue (D/A) conversion provides a possibility since the quantization error is a signal that can be used to

excite the identified process. Therefore, the aim of this thesis is to find the dynamics of a controlled system from closed-loop quantized data.

1.2 Definitions and concepts

This section will first give some basic definitions on system identification. Closed-loop identification will then be discussed. Emphasis will be given to two specific closed-loop identification methods — closed-loop identification with external excitation and relay-based identification.

1.2.1 Basic definitions

The focus of the work is the determination of the behavior of controlled system, while it remains under closed loop control. The Instrument Society of America gives the following definitions.

> Controlled system: The collective functions performed in and by the equipment in which the variable(s) is (are) to be controlled.
>
> (ANSI/ISA-S51.1-1979:29)

> Closed loop: A signal path, which includes a forward path, a feedback path and a summing point, and forms a closed circuit.
>
> (ANSI/ISA-S51.1-1979:8)

> Feedback control: Control in which a measured variable is compared to its desired value to produce an actuating error signal which is acted upon in such a way as to reduce the magnitude of the error.
>
> (ANSI/ISA-S51.1-1979:11)

For the hot tank example, the hot water tank is a controlled system because it is desired to maintain the hot water at a desired temperature by means of a heating element. The tank and the heater comprise a closed loop.

1.2.2 System identification definitions

When a new system is encountered, some concept of how its variables relate to each other is needed. With a broad definition, such an assumed relationship among observed signals is a *model* of the system. There are many different kinds of models; for example, linear systems can

be described by their impulse or step responses. These are *graphical models*. In more advanced applications, it may be necessary to use models that describe the relationships among the system variables in terms of mathematical expressions like difference or differential equations. These are *mathematical (or analytical) models*. Mathematical models may be further classified into different types, such as continuous time or discrete time, lumped or distributed, linear or nonlinear. Throughout this thesis, only mathematical models of linear lumped systems will be discussed.

This subsection now introduces some basic concepts that will be valuable when describing and analyzing identification methods (see, for example, Ljung, 1987; Söderström and Stoica, 1989).

Process and system:

> The physical reality that provides the experimental data will be generally referred to as the *process*. ... The word *system* denotes a mathematical description of the true process.
>
> (Söderström and Stoica, 1989:9)

In Figure 1-1, the true process is G_0. For instance, if expressed as continuous linear system, it may take the form of

$$G(s) = \frac{Ke^{-Ls}}{Ts+1} \qquad (1.1)$$

where $G(s)$ is the transfer function, K is the process steady-state gain, L is the dead time, T is the time constant and s is *Laplace variable*.

System identification:

> *System identification* is the field of modeling dynamic systems from experimental data.
>
> (Söderström and Stoica, 1989:1)

In Figure 1-1, a batch of data may be collected.

$$Z^N = [y(1), u(1), ..., y(N), u(N)] \qquad (1.2)$$

Where u is the input and y is the output. The task of system identification is to determine an estimate $\hat{G}$ of the true process G_0 from the data set Z^N. Still for the hot water tank in a domestic house, the behavior to be identified is the relationship between tank temperature and heater input, e.g. how much heat is needed to raise the temperature by 1°C (the gain) and how long it takes for the tank to attain some fraction of the desired temperature (63% is commonly used).

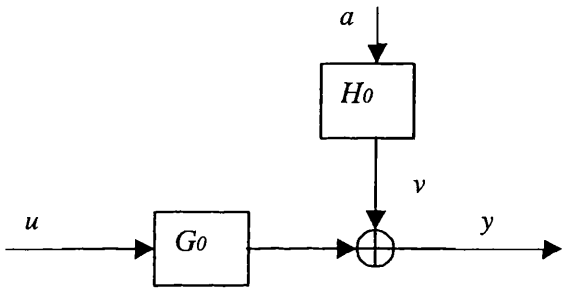


Figure 1-1 *Open-loop system in which G_0 is the true process and H_0 is the noise model; a is white noise and v is the noise applied on the process.*

Nonparametric methods: A model that is described by a curve, function or table is called a *nonparametric model.* A step response, an impulse response and a frequency diagram such as Bode and Nyquist plots are examples of nonparametric models. Methods that aim at determining nonparametric models by direct techniques without first selecting a confined set of possible models, are often called *nonparametric methods* since they do not (explicitly) employ a finite-dimensional parameter vector in the search for a best description.

In system identification, commonly used nonparametric methods include, for example, the correlation analysis method and the spectral analysis method (Ljung, 1987).

Model Structure $\mathcal{M}$:

In many cases, it is relevant to deal with *parametric models.* Such a model is characterized by parameter vector θ. The corresponding

> model will be denoted $\mathcal{M}(\theta)$. When θ is varied over some set of feasible values, we obtain a model set (a set of models) or a *model structure* $\mathcal{M}$.
>
> (Söderström and Stoica, 1989:9)

A general parametric model structure (Ljung, 1991) is

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t-nk) + \frac{C(q)}{D(q)}e(t) \tag{1.3}$$

where A,B,C,D and F are polynomials in the delay operator q^{-1}:

$$A(q) = 1 + a_1 q^{-1} + \cdots\cdots + a_{n_a} q^{-n_a} \tag{1.4}$$

$$B(q) = b_1 q^{-1} + \cdots\cdots + b_{n_b} q^{-n_b} \tag{1.5}$$

$$C(q) = 1 + c_1 q^{-1} + \cdots\cdots + c_{n_c} q^{-n_c} \tag{1.6}$$

$$D(q) = 1 + d_1 q^{-1} + \cdots\cdots + d_{n_d} q^{-n_d} \tag{1.7}$$

$$F(q) = 1 + f_1 q^{-1} + \cdots\cdots + f_{n_f} q^{-n_f} \tag{1.8}$$

in which, n_a, n_b, n_c, n_d and n_f are the orders of the polynomials A,B,C,D and F, k is the number of delays from input to output.

Other linear parametric models are special cases of the above general parametric model structure. For example, the general parametric model structure becomes an ARMAX model when both n_d and n_f equal to zero.

$$A(q)y(t) = B(q)u(t-nk) + C(q)e(t) \tag{1.9}$$

The general parametric model structure becomes *Box-Jenkins* (BJ) model when n_a equals to zero.

$$y(t) = \frac{B(q)}{F(q)}u(t-nk) + \frac{C(q)}{D(q)}e(t) \qquad (1.10)$$

ARX model is another special case of the general parametric model structure when n_c, n_d and n_f all equal to zero.

$$A(q)y(t) = B(q)u(t-nk) + e(t) \qquad (1.11)$$

In the special case of ARX model where $n_a = 0$, ARX model becomes FIR (finite impulse response) model.

$$y(t) = B(q)u(t-nk) + e(t) \qquad (1.12)$$

Parameter estimation methods: When a set of candidate models has been selected and it is parametrized as a model structure using a parameter vector θ, the search for the best model within the set then becomes a problem of determining or estimating θ. If the data set Z^N has been collected, then the definition is as follows:

> The problem we are faced with is to decide upon how to use the information contained in Z^N to select a proper value $\hat{\theta}_N$ of the parameter vector, and hence a proper member $\mathcal{M}(\hat{\theta}_N)$ in the set $\mathcal{M}^*$. Formally speaking, we have to determine a mapping from the data Z^N to the set D_m (set of values over which θ ranges in a model structure $\mathcal{M}$)
>
> $$Z^N \rightarrow \hat{\theta}_N \in D_m \qquad (1.13)$$
>
> Such a mapping is *parameter estimation method.*
>
> (Ljung, 1987:170)

In system identification, commonly used parametric methods include, for example, the least-squares (LS) method and the maximum likelihood (ML) method.

Bias: When the expected value of an estimate $\hat{\theta}$ deviates from the true value θ_0, i.e. $E\hat{\theta} \neq \theta_0$, where E is the expectation operation, the estimate $\hat{\theta}$ is biased. The difference $E\hat{\theta} - \theta_0$ is called the *bias*. Otherwise, if $E\hat{\theta} = \theta_0$, $\hat{\theta}$ is said to be *unbiased.* (Söderström and Stoica, 1989:18)

Consistency: An estimate $\hat{\theta}$ is consistent if $\hat{\theta} \to \theta_0$ as $N \to \infty$, where N is the number of data points used in the estimate. A reasonable alternative is 'limit with probability one' (Söderström and Stoica, 1989:19).

Identifiability: This is a concept that is central in identification problems. Loosely speaking, identifiability is whether the identification procedure will yield a unique value of the parameter θ and whether the resulting model is equal to the true system. The issue involves aspects on whether the data set Z^N is informative enough to distinguish between different models as well as whether different values of θ can give equal models (Ljung, 1987:99). Generally, the term system identifiability refers to the joint identifiability of the process G_0 and the disturbance/noise model H_0 (as shown in Figure 1-1). However, throughout this thesis, only the identifiability of the process G_0 is concerned because only the true process is of interest.

Persistent exciting: This is an important concept which will be used repeatedly in this thesis.

> A quasi-stationary signal $\{u(t)\}$ with spectrum $\phi_u(\omega)$ is said to be persistently exciting if
>
> $$\phi_u(\omega) > 0, \qquad \forall \omega \qquad (1.14)$$
>
> (Ljung, 1987:364)

1.2.3 Closed-loop identification (CLID)

Performing identification experiments in closed loop may be necessary due to safety or economic reasons or because of inherent feedback (Ljung, 1987).

> Closed-loop Identification is the procedure of collecting experimental data from a process while it is operating under feedback control and then using those data to develop a dynamical model for the system.
>
> (MacGregor and Fogal, 1995:163)

A large number of academic papers on the topic of closed-loop identification (CLID) have been published. There are three main categories of closed-loop identification.

- The first category is the general identification technique for stochastic processes. The process input is usually an independent, persistently exciting signal (i.e. a broad-band excitation signal that excites the process at many frequencies). The identification result is usually a discrete-time model, for example (Ljung, 1987)

$$G(q) = q^{-k}\frac{B(q)}{A(q)} \tag{1.14}$$

where $G(q)$ is pulse transfer function, B and A are polynomials, q is *forward shift operator*, correspondingly q^{-1} is *backward shift operator* and k is the number of delays from input to output.

This identification result can then be used for model-based control design. The typical overview papers on this topic are Gustavsson *et al.* (1977) and Forssell and Ljung (1999).

- The second category occurs in the context of auto-tuning. It is an alternative to the Ziegler-Nichols frequency response method. A nonlinear relay is connected to replace the available controller. No input signal is used in this method. The final identification result is usually a continuous model, or transfer function, for example (Åström and Hägglund, 1995)

$$G(s) = \frac{Ke^{-Ls}}{Ts+1} \tag{1.15}$$

where $G(s)$ is the transfer function, K is process steady-state gain, L is the dead time, T is the time constant and s is *Laplace variable*.

This continuous model is then used for Proportional-Integral-Derivative (PID) controller tuning. The well-known work is Åström and Hägglund (1984a). This is called Relay-based identification.

- In the third category, a step or pulse is used as process input signal. The end product is also a continuous model, which is consequently used for PID controller tuning. For example, Yuwana and Seborg (1982) proposed the estimation of the critical point of the process (i.e. where its frequency response has a phase shift of 180°) based on a first-order plus dead-time (FOPDT) model identified from a step set-point response of the proportional control system. This is called a proportional (P) controller method.

In the literature, when closed-loop identification is referred to, it always means the first category by default. For easy notation so as to avoid confusion, the first category is called *closed-loop identification with external excitation* in this thesis since an independent persistently exciting signal (i.e. a broad-band excitation signal that excites the process at many frequencies) is used to guarantee closed-loop identifiability in all the methods of this category. The second category is called *relay-based identification.* In this thesis, only the first two categories will be studied in detail since they will be further developed to formulate a new closed-loop identification scheme.

1.2.4 Closed-loop identification with external excitation

External excitation is a dither signal injected into the original closed-loop system to excite the process for the purpose of system identification. A single-input single-output (SISO) closed-loop system (adopted from MacGregor and Fogal, 1995) illustrated the material to be presented in this subsection. It provides a general structure for CLID with external excitation. The excitation is a dither signal introduced at *d.*

In Figure 1-2, $G_0(q)$ represents the true process, and the disturbance $v(t) = H_0(q)a(t)$ represents the effect of all unmeasured process disturbances on the measured output $y(t)$, q is the forward shift operator. $C(q)$ is the feedback controller, and the set-point $y_{sp}(t)$ and the 'dither' signal *d(t)* are input signals that may be injected to aid in the identification. In this thesis, only $y_{sp}(t) = 0$ is considered, that is to say, only *d(t)* is used as an external input signal.

The identification task is to use *y(t)* and *d(t)* to identify the process. This can be done in a number of ways, for example, Huang and Shah (1997), Van den Hof and Schrama (1993).

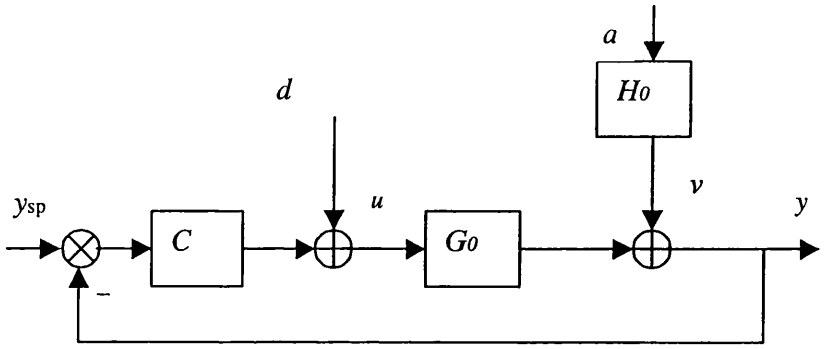


Figure 1-2 *Closed-loop system*

1.2.5 Relay-based identification

In normal condition, the process works under the control of a controller (Figure 1-3). The idea was to introduce a nonlinear feedback of the ideal relay in order to generate a sustained oscillation during the identification. The system then starts to oscillate. The period and the amplitude of the oscillation are determined when steady-state oscillation is obtained (Åström and Hägglund, 1995). The measured oscillation information can be used in two ways, either to estimate the critical point of the process, that is where its frequency response has a phase shift of 180° and thus may be regarded as an automated Ziegler-Nichols test.

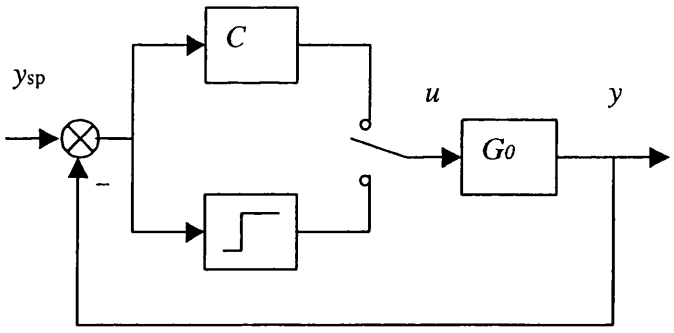


Figure 1-3 *Standard relay feedback system*

1.3 Problem analysis and key issues

1.3.1 Problem analysis

At this moment, a natural question raised often is that: With so many choices for closed-loop identification methods, why are new schemes for closed-loop identification still needed?

Within the first category, CLID methods rely on external perturbations to excite the process. These perturbations, such as adding a dithering signal to the control loop, are not a part of the closed-loop and are external to the process under feedback control. The role of these signals is to excite the process for the purpose of identifying the true dynamics of the process while the process is operating under feedback control. This excitation has the unfortunate effect of degrading the control performance because it is an external disruption to the process. It cannot be avoided, but should be minimized.

On the other hand, as described in subsection 1.2.5, information on the critical point can be identified from relay feedback tests. These parameters allow PID tuning settings to be calculated, for instance from Ziegler-Nichols rules or other standard tuning methods. Relay tuning has been successful. For example, the Alfa Laval Automation ECA400 controller that is commercially available is based on relay identification (Åström and Hägglund, 1995). However, the problem is that replacement of the PID controller by the relay disrupts the operation of the control loop.

Achieving identification during normal operations without external excitation or disruption is an ideal target. But theoretical analysis and simulation studies for closed-loop identifiability in section 3.1 will show that it is impossible. So the best solution should identify the process dynamics with as small disruption to the controlled process as possible.

The observation that quantization caused by analogue to digital (A/D) conversion and/or digital to analogue (D/A) conversion in the process instruments injects an excitation into a single-input-single-output (SISO) loop even during normal process operations should be inspiring (EPSRC Grant Proposal GR/R06687 by Thornhill *et al.* 2000).

It is assumed that the data are measured with a precision of 10 bits (which is quite reasonable for typical A/D converters). Figure 1-4 shows a SISO control loop with 10-bit A/D converter. The *quantization error* (the quantized observation minus the unquantized underlying signal, i.e.

$y - y_a$ in Figure 1-4) acts as the excitation. The 10-bit A/D converter for measurement purpose is not very flexible. Therefore, the last scheme should be like that displayed in Figure 1-5 with another A/D converter inserted in the feedback path. The excitation can be adjusted in two ways, either by increasing the quantization interval of the additional A/D converter, or by altering the quantization sampling rate (i.e. the times at which the quantized samples are taken) of the additional A/D converter. The quantization error ($y_q - y$ in Figure 1-5) will be used as the excitation required by closed loop identification.

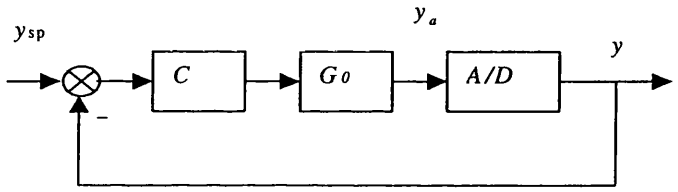


Figure 1-4 *Closed-loop system with 10-bit A/D converter*

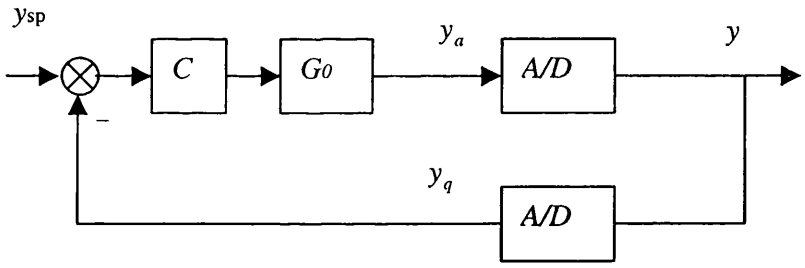


Figure 1-5 *Closed-loop identification using quantized data*

It is the aim of this thesis to find novel ways to meet the challenge of achieving identification during normal operations with minimum disturbance. Quantizer signal processing seems to be a potential solution. Therefore, the motivation of the work is to find how to use the quantizer to excite sufficient information for system identification while minimizing the disturbance.

1.3.2 Key issues

The aim of this thesis is the identification of the dynamics of a controlled system from closed-loop data while causing minimum

disruption to the controlled process. The key idea is to exploit measurement quantization errors caused by analogue to digital conversion as an excitation signal.

To achieve the target described above, the following key issues need to be addressed.

- To explore the different ways that can meet closed-loop identifiability. This will provide theoretical fundamentals why closed-loop identification can be achieved based on quantization.
- To compare the advantages and disadvantages of the CLID with external excitation method and the relay-based identification.
- To recover an underlying signal from quantized measurements. The quantization error is the quantized observation minus the unquantized/underlying signal. After the quantized data has been obtained, the key issue is to implement an algorithm to recover the underlying signal.
- To explore the relationship between quantization interval and the information content of quantization error. This information content includes power and bandwidth.
- To select the optimum quantization interval or range for different closed-loop identification methods.
- To formulate new CLID scheme for controlled system identification using quantized data.
- To demonstrate accurate system identification using quantized plant data with pilot plant experimentation

1.4 Outline of approaches

1.4.1 Case study

Case studies will be extensively used to validate theory and to gain insights for different closed-loop identification methods. In section 3.1, case studies will be used to validate the theory on closed-loop identifiability. Then in section 3.2, case studies will be used to duplicate the two-stage method (Van den Hof and Schrama, 1993) and the two-step method (Huang and Shah, 1997). In the same way, the relay identification methods (Li. *et al.*, 1991; Wang, QG *et al.*, 1997a and Wang, QG *et al.*, 1997b) have been duplicated in section 3.3. The noise influence in the relay-based identification has been explored in section 3.3. The two categories of closed-loop identification methods will also be compared by case study in section 3.4.

1.4.2 Simulation

Simulation is widely used to test the theory and the algorithms. In Chapter 3, all case studies will be implemented through simulation. In Chapter 5 and Chapter 6, transfer function model simulation and pilot-plant model simulation will be used to demonstrate the closed-loop identification scheme based on quantization. In Chapter 7, simulation will be used to demonstrate the superiority of Quantized Regression (QR) algorithm over other methods (e.g. the Linear Least Square (LLS) method) in reconstructing from quantized measurements.

All the simulation analyses were performed using MATLAB version 6.0 together with Simulink version 4.0 and System Identification Toolbox version 5.0 (The MathWorks; Natwick, MA).

1.4.3 Experimentation

Experimentation is a convincing way to test the theory and the algorithms. In Chapter 5 and Chapter 6, experimentation with a pilot plant in the Computer Process Control group in the University of Alberta will be used to demonstrate the closed-loop identification scheme based on quantization. In Chapter 7, experimentation data of the pH control in a buffered fed-batch yeast fermentation process will be used to demonstrate the superiority of quantized regression (QR) algorithm over other methods (for example, the linear least square (LLS) method) in recovering the underlying signal from quantized measurements.

1.5 Outline of the thesis

In this thesis, the general objective is described in Chapter 1 on the basis of introducing the important definitions and concepts as well as two main categories of closed-loop identification methods: closed-loop identification with external excitation and relay-based identification. Emphases are put on the problem analysis and key issue to be addressed. Chapter 2 is a literature review in which the developments of the two categories of identification methods and the relevant closed-loop identification based on quantization are evaluated. The topic of closed-loop performance assessment (CLPA) is also reviewed since the concept and method will be further used in Chapter 5 and Chapter 6.

The closed-loop identifiability theory is analyzed in Chapter 3 together with a case study. Advantages and disadvantages of the two methods can be inferred from the case studies

presented in Chapter 3. The main advantage for relay-based identification is that the sustained oscillation can be implemented automatically. The main disadvantage for relay-based identification is that the controller must be replaced by an ideal relay, which means disruption to the control system. The main advantage for closed-loop identification with external excitation is that over a range of frequencies information of the process can be identified and the identified model is of higher accuracy. The main disadvantage for closed-loop identification with external excitation is that *a priori* information on the critical frequency of the identified system is needed for the design of the persistently exciting signal. These case studies will also provide a benchmark for further work of assessing the new closed-loop identification schemes.

Chapter 4 starts with the definition of quantizers. The relationship between quantization interval and the information content (power and bandwidth) of the quantization error is then presented. A new closed-loop identification scheme is proposed – when quantization interval ranges from small to large, the closed-loop identification with external excitation method and the relay-based identification method can be applied correspondingly. The stability influence of the newly inserted quantizer to the closed-loop system is also analyzed in Chapter 4. Chapter 5 introduces the computation methods such as a nonlinear test method, model accuracy measure, the simulation method and the experimentation method. All the results and discussions are written in detail in Chapter 6.

Chapter 7 is a self-contained chapter, which presents the Quantized Regression (QR) algorithm used to recover the underlying signal from quantized measurements. The QR algorithm was implemented successfully but was not in the end used in the closed-loop identification procedure because its large computation makes it difficult to use on-line.

The main novel contributions of this thesis are summed up in Chapter 8. In the same chapter, further suggestions on the future research along this method are also presented.

CHAPTER 2

LITERATURE REVIEW

This chapter focuses on closed-loop identification with external excitation (Ljung, 1987) and relay-based identification (Åström and Hägglund, 1984a and 1984b). It also includes some previous research on closed-loop identification with a quantizer. The relevant topic of closed-loop performance assessment (Harris, 1989) is also discussed since it will be used later in Chapter 4, Chapter 5 and Chapter 6 of this thesis.

2.1 Closed-loop identification with external excitation

In this section, a general structure for closed-loop identification with external excitation will be described first. The efforts will then be directed to classification and development of the CLID approaches and methods. The recent development of closed-loop identification without external excitation is also discussed. The conclusion is given in the end that some kind of external excitation is necessary for closed-loop identification.

2.1.1 Introduction

The single-input single-output (SISO) closed-loop system (adapted from Ljung, 1987) illustrates the material to be presented in this section. It provides a general structure for CLID with external excitation. The excitation is a dither signal introduced at *d*.

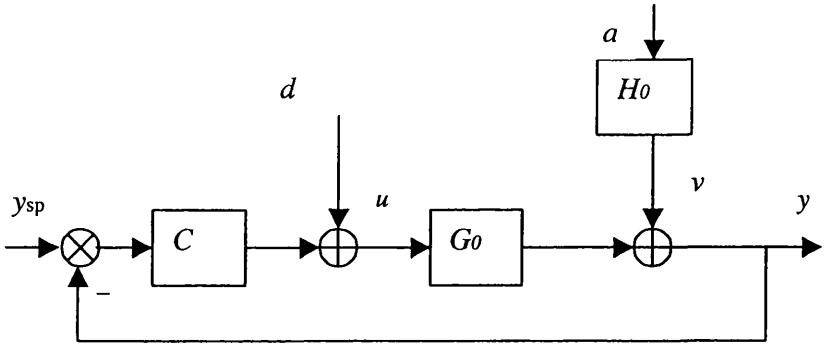


Figure 2-1 *Closed-loop system*

The true system is assumed to be

$$y(t) = G_0(q)u(t) + H_0(q)a(t) \qquad (2.1)$$

where $G_0(q)$ represents the true process, and the disturbance $v(t) = H_0(q)a(t)$ represents the effect of all unmeasured process disturbances on the measured output y(t), q is the forward shift operator. The feedback loop is

$$u(t) = C(q)(y_{sp}(t) - y(t)) + d(t) \qquad (2.2)$$

where $C(q)$ is the feedback controller, and the set-point $y_{sp}(t)$ and the 'dither' signal $d(t)$ are input signals that may be injected to aid in the identification. As stated in subsection 1.2.4, only $y_{sp}(t) = 0$ is considered in this thesis. Thus, the task of identification is how to use the available information (such as $d(t)$, $u(t)$ and $y(t)$) to get an estimate $\hat{G}$ of the true process G_0.

2.1.2 Motivation

Closed-loop identification (CLID) identifies a process model while the process is operating under feedback control. Closed-loop identification might be necessary because the system is unstable in open loop or the system contains inherent feedback mechanisms (Ljung, 1987).

Gustavsson *et al.* (1977) gave an overview of CLID mainly on closed loop identifiability and accuracy. Their conclusions are that prediction error methods can be applied to identify linear model using data collected in a closed-loop and identification is possible in some cases, for example, external excitation, setpoint change, time-varying controller or nonlinear controller. Forssell and Ljung (1999) revisited CLID under the prediction error framework. Van den Hof and Schrama (1995) systematically reviewed the problem of control-relevant CLID. Their main conclusion is that iterative closed-loop identification and control design can improve performance.

2.1.3 Framework of approaches and methods

Different assumptions of feedback configurations lead to a classification of CLID approaches (Gustavsson *et al.*, 1977; Forssell and Ljung, 1999):

(1) The Direct Approach: the feedback is ignored and the controlled system is identified using measurements of the input $u(t)$ and the output $y(t)$ (see Figure 2-1).
(2) The Indirect Approach: the closed-loop transfer function $Y(s)/D(s)$ is identified at first and the open-loop parameters are then determined using the knowledge of the controller $C(q)$.

(3) The Joint Input-output Approach: the input *u(t)* and the output *y(t)* are jointly viewed as the output from a system driven by the external excitation signal *d(t)* and noise *a(t)*. Some method is used to determine the open-loop parameters from an estimate of this augmented system. It is not required to know the controller *C(q)*, but the controller must be known to have a certain structure. In detail, three branches exist inside Joint Input-Output Approach (Gustavsson *et al.*, 1977; Forssell and Ljung, 1999). They are: the Coprime Factor method, the Two-stage (step) method, and the Projection method.

Thus, the framework for CLID with External Excitation is:

- Direct approach
- Indirect approach
- Joint input-output approach
    - Coprime factor method

      Coprime factors means having no common factors (Schrama, 1991).
    - Two-stage (step) method

      There are two stages or steps; the first stage/step is to identify sensitivity function and the second stage/step is to identify the process model (Van den Hof and Schrama, 1993; Huang and Shah, 1997).
    - Projection method

      This method is similar to two-stage/step method, the difference is a non-causal FIR filter used in the first sage/step; the first step in the method can be viewed a least squares projection of the process input *u(t)* onto external excitation *d(t)* – hence it is called projection method (Forssell and Ljung, 1998; Forssell and Ljung, 1999).

On the other hand, in the context of open-loop system identification, the textbook of Ljung (1987) classifies open-loop system identification into

- Parametric method

  With this method, a model structure (finite dimensional parametric vector) has to be chosen at first, followed by parameter estimate and model validation.
    - Prediction Error family

      Prediction error means the measurements minus the corresponding one-step-ahead predictions derived from the model; a general term *prediction error identification methods* is also used to represent this family of methods (Ljung, 1987, pp170-171).

- Least square method, in which the estimate is obtained by choosing the criterion as minimizing the sum of the squared one-step-ahead prediction errors, it is also called *linear regressions* in the literature; it is a special case of *prediction error identification methods* (Ljung, 1987: 175-176).
- Instrumental Variable method: the key difference between IV methods and LS methods is that LS methods can only be used when the measurements are contaminated by white noise and IV methods can be used under coloured noise for consistent estimation (Söderström and Stoica, 1989, pp187, pp261; Ljung, 1987, pp192-193).

- Subspace family

  Subspace methods estimate a state space model of a multivariable process directly from input/output data. The main part of a subspace method consists of matrix singular value decomposition (SVD) and linear least-squares estimation which are numerically simple and reliable (Zhu and Butoyi, 2002).

- Non-parametric method

  A model is described by a curve (for example, impulse response, step response), a function or a table; a finite-dimensional parameter vector is not explicitly employed in the search for a best description (Ljung, 1987, pp141).

  - Correlation method in which the input is white (if it is not white, a whitening filter can be used to make it white), the impulse response can be calculated through correlation analysis (i.e. the cross covariance function between the input and the output) (Ljung, 1991, pp1-17).
  - Spectral analysis method in which the frequency response is obtained by dividing the cross spectrum of the output and the input with the spectrum of the input (Ljung, 1991, pp1-18).

The above two classifications complement each other. Within each of these approaches (direct, indirect, joint input-output approach), different methods (such as: prediction error, subspace) can be used. It should be noted that the Direct approach is only applicable with the prediction error method and some of the subspace methods. The reason for this is the unavoidable correlation between the unmeasurable noise and the input. With the indirect and joint input-output approaches, the closed-loop problem is converted into an open-loop one since the external input signal is uncorrelated with the noise. Thus all open loop methods can be used within the indirect and joint input-output approaches.

2.1.4 Specific approaches and methods

In this subsection, different methods will be placed in the context of the approach used. The aim is to make the classification in section 2.1.3 more concrete to present the methods that have potential use in the thesis later.

Direct approach

The direct approach can only be used with the prediction error method and some of the subspace methods. The reason for this is the unavoidable correlation between the unmeasurable noise and the input.

Indirect approach

Theoretical analysis reveals that application of the conventional least-squares (LS) method in indirect identification can always result in inconsistent parameter estimates of the closed-loop system as well as of the open-loop process in the presence of coloured noise. Zheng and Feng (1995) presented an LS type method to cope with the bias problem in indirect identification of closed-loop systems subjected to coloured disturbances. The proposed method is based on the bias-correction principle. The suggested method can yield consistent parameter estimates. Generally, this kind of method is referred to as bias-free least squares (BFLS).

The instrumental variable (IV) methods form a different class of identification methods that is related, but not equivalent, to the least square (LS) methods inside the prediction error framework. The key differences between IV methods and LS methods are that LS methods can only be used when the measurements are contaminated by white noise and IV methods can be used under coloured noise for consistent estimation (Söderström and Stoica, 1989, pp187, pp261). The closed-loop IV methods have been studied by Söderström *et al.* (1987). Under weak assumptions, the estimates are consistent and asymptotically Gaussian distributed. To guarantee the identifiability of the closed-loop, a measurable external signal such as a reference or a setpoint is needed.

Zhang *et al.* (1997) studied the connection between the Bias-free Least Squares (BFLS) method and the Instrumental Variable (IV) family and showed that the BFLS method is a kind of weighted instrumental variable method whose results do not depend on the order of the controller.

Joint input-output approach

Schrama (1991) proposed a framework for open-loop identification of the coprime factors of the unknown plant in dealing with the approximate closed-loop identification problem. This framework allows the feedback-controlled plant to be identified by the application of any open-loop identification method.

Van den Hof and Schrama (1993) introduced a two-stage method. The transfer function of a linear plant can be consistently estimated on the basis of data collected from closed loop experiments, even in the situation where the model of the noise disturbance is not accurate. Huang and Shah (1997) modified the two-stage method described above and demonstrated better accuracy. Both researches require that a persistently exciting external signal (i.e. a broad-band excitation signal that excites the process at many frequencies). Esmaili *et al.* (2000) discussed the asymptotic and finite data behaviour of some closed-loop identification methods including the two-step method (Huang and Shah, 1997).

Forssell and Ljung (1998) initiated a projection method, which comprises the same two steps as the two-stage or two-step method. The only difference is that in the first step one should use a non-causal FIR model instead of a causal high-order FIR or ARX model.

ARX model is a commonly used parametric model, for example

$$G(q) = q^{-k}\frac{B(q)}{A(q)}\ ; \qquad H(q) = \frac{1}{A(q)} \tag{2.3}$$

where B and A are polynomials in the delay operator q^{-1}:

$$A(q) = 1 + a_1 q^{-1} + \cdots\cdots + a_{n_a} q^{-n_a} \tag{2.4}$$

$$B(q) = 1 + b_1 q^{-1} + \cdots\cdots + b_{n_b} q^{-n_b} \tag{2.5}$$

in which, n_a and n_b are the orders of the polynomials A and B, k is the number of delays from input to output. In the special case where $n_a = 0$, the ARX model becomes an FIR (finite impulse response) model. Non-causal FIR model means that

$$u(t) = S(q,\beta)d(t) + e(t) = \sum_{k=-M1}^{M2} s_k d(t-k) + e(t) \qquad (2.6)$$

where $M1$ and $M2$ are positive integers and d is the input. The main feature is that the sum contains both the positive items and the negative items - which means past and future.

2.1.5 Closed-loop identifiability, accuracy and efficiency

The issues of identifiability, accuracy and efficiency were important topics in the early stages of closed-loop identification in 1960s and 1970s. They remain under investigation in the literature.

Closed-loop identifiability

Gustavsson *et al.* (1977) is the most widely referred paper on closed-loop identifiability. MacGregor and Fogal (1995) discussed the closed-loop identifiability with simulation examples. They showed that the nonparametric methods yield no information on the true system if data from purely feedback operation is used. Under the same condition, the parametric methods can identify the true system only when the fairly strict necessary and sufficient conditions on the true process and controller orders are satisfied (i.e. order of the controller should be greater than, at least equal to, the order of the process). In Gustavsson *et al.* (1977), this is called system identifiability (SI).

It is apparent that for closed-loop identification something must be done to break the dependency between the input signal and the process disturbance. This can be accomplished in either of the two ways (MacGregor and Fogal, 1995; Bartee and McFarlane, 1998).
(1) By injecting an independent, persistently exciting signal into the feedback loop.
(2) By switching between two or more feedback controllers.
Either of these will guarantee that necessary and sufficient conditions for identifiability be satisfied. This is known as strong system identifiabilty (SSI) (Gustavsson *et al.*, 1977).

The third way to break the dependency between the input signal and the process disturbance is to insert nonlinearity (page 366 of Ljung, 1987). For instance, switching between two or more feedback controllers is one way to implement nonlinearity.

SSI infers the following properties (Gustavsson *et al.*, 1977):
(1)Closed-loop data can be used as though it were open-loop data for model identification.

(2)All statistical estimation methods (non-parametric method and parametric method) and validation techniques are valid.

(3)Descriptive transfer function models can be identified, even when they are over-parameterized.

Therefore injecting external excitation into a system makes it to be SSI. The injection of an independent signal into the loop provides the user with a greater flexibility for designing the identification experiment to achieve the desired objectives. That is why this category is called closed-loop identification with external excitation.

Accuracy

The problem of assessing the quality of the estimated model, i.e. the accuracy, is an important issue in system identification (Gustavsson *et al.*, 1977). Model-plant error consists of two parts. One component is due to the fact that the true system is not within the model structure that has been chosen (e.g. under-modeling), and this is called the *bias* error. The second component of the total error is due to noise corruption of the observed data (disturbance effects or the random error). This is termed *variance* error.

The direct approach in which the feedback is ignored typically gives better accuracy than the indirect approach and joint input-output approach where the closed-loop problem is converted into an open-loop one. Ljung and Forssell (1997) derived variance results for a number of closed-loop identification methods using standard prediction error theory. By studying the asymptotic variance for the parameter vector estimates, they showed that indirect methods fail to give better accuracy than the direct method. So the direct approach should be regarded as the first choice of methods for closed-loop identification whenever possible. Gevers *et al* (2001) derived the asymptotic variance expressions for models that are identified on the basis of closed-loop data. The CLID methods considered include the direct method, as well as indirect methods such as coprime factorization method, dual Youla/Kucera parameterizations and the two-stage method by Van den Hof and Schrama (1993). They showed that different methods (direct method and indirect method) lead to the same asymptotic variance.

Even so, under the prediction error framework, using the direct approach, the consistency (no bias) of the model estimate essentially relies on the correct modeling of the system disturbance dynamics (Forssell and Chou, 1998). MacGregor and Fogal(1995) have shown the same result with simulation examples. However, in the indirect approach and joint input-output approach,

the consistency of the resulting model can be achieved without putting too much attention on the system disturbance model. For example, with the two-stage method proposed by Van den Hof and Schrama (1993), the transfer function of a linear plant can be consistently estimated even when the model of the noise disturbance is not accurate.

Ljung and Forssell (1999) studied the same problem i.e., direct identification of systems operating in closed-loop gives biased results whenever the true noise characteristics are not correctly modeled. When signal-to-noise ratio is large, the bias error will be small. But the only way to completely remove the bias is to use the indirect method.

Under certain circumstances, the instrumental variable (IV) method can give the same level of accuracy as the direct PE method. In general, though, the accuracy will be sub-optimal (Forssell and Chou, 1998). With optimally chosen instruments, the accuracy of IV method coincides with that of the direct method. However, this requires exact knowledge of the noise model and therefore cannot be achieved in practice.

Bias-free least squares (BFLS) method proposed by Zheng and Feng (1995) to cope with the bias problem in indirect identification of closed-loop systems subjected to coloured disturbances is based on the bias-correction principle. This might be useful in designing new closed-loop identification scheme with quantization signal processing.

Efficiency

> An unbiased estimate is said to be efficient if its covariance equals the Cramer-Rao bound.
>
> (Norton, 1986:117)

Efficiency is defined to be Cramer-Rao bound divided by the estimation variance. The main practical significance of efficiency lies in determining whether further efforts needed to improve the estimate. Forssell and Chou (1998) studied the efficiency of a number of closed-loop identification methods, mainly prediction error (PE) methods and instrumental variable (IV) methods. The reason why direct applied prediction error method gives better accuracy is that in direct approach all the input signal power is utilized in reducing the variance. In other approaches, only certain part of the input spectrum (i.e. the noise-free part) is used, the signal-to-noise ratio is reduced and the variance is increased.

2.1.6 Joint identification and control

From the 1990's, control relevant identification (or joint identification and control) has attracted much attention. The goal is to construct models that are suitable for control design. The key idea in the joint identification and control strategy is to identify and control with the objective of minimizing a joint global control performance criterion. Van den Hof and Schrama (1995) reviewed the problem of joint identification and control. They discussed attempts to provide a separate analysis of both steps (approximate identification and model-based control design), and to accomplish a joint (robust) performance criterion of both parts. The conclusion is that the latter can have better performance.

The study of control-relevant identification by Hjalmarsson *et al*, 1996 showed that the best identification strategy is to identify the process under feedback with the intended controller in use. Callafon (1998) studied the integration of closed loop identification with robust control design and the application of this method to manufacturing of a wafer stage.

To perform dual roles of identification and control simultaneously within a constrained Model Predictive Control (MPC) framework is termed MPCI. Nikolaou and Eker (1998) introduced a new MPCI variant. Process inputs are constrainted to excite the process as much as possible for the generation of maximum parameter information. Gopaluni *et al.* (2002a and 2002b) also studied closed-loop identification with MPC controller as the intended controller. The conclusion is that it is important to minimize multi-step-ahead predictions, as opposed to one-step-ahead prediction errors, if MPC Controller is used. Zhu and Butoyi (2002) studied multivariable and closed-loop identification of industrial processes for use in MPC. Twu case studies were used to demonstrate the advantages of closed-loop identification.

2.1.7 The problem of CLID without excternal excitation

As described above, the paper of Gustavsson *et al.* (1977) showed that the injection of an independent signal into the loop guarantees the closed-loop identifiability. Generally, CLID methods rely on external perturbations to excite the process.

As analyzed in subsection 1.3.1, on the one hand, excitation signal has to be injected to excite the process for the purpose of identifying the true dynamics of the process while the process is operating under feedback control. On the other hand, the exciting signal is not a part of the closed-loop and is considered external to the process under feedback control. Therefore, this

excitation has the unfortunate effect of degrading the control performance and should be minimized in some way.

Due to the above reason, researchers have been motivated by the desire to use only routine closed-loop operating data (i.e. without external excitation) for model identification. Sun *et al.* (2001) proposed a new identification algorithm for a linear discrete-time closed-loop system. Identifiability condition was not satisfied because neither an external excitation was injected nor the controller was switched. The algorithm is based on output over-sampling scheme (i.e. output sampling interval is much smaller than input sampling interval). The over-sampled output data contain more information about the system structure. The restriction of the proposed method is that the model structure (including the order and delay time) of the plant is assumed to be known. Bartee and McFarlane (1998) studied the identification of process models using routine operating data (or archived data) with simulation examples in a fluid catalytic cracking unit. They showed that with routine data (from systems satisfying SI), only when the true system structure is known, can an adequate model be identified. These publications showed that routine or archived data have limited practical use in closed-loop system identification. Therefore, there is a need for some kind of external excitation.

2.1.8 Conclusions

From the discussions in this section, the conclusions are:

- Closed-loop identifiability (MacGregor and Fogal, 1995) is an important topic.
- Van den Hof and Schrama (1993) and Huang and Shah (1997) are all two-stage (step) methods of joint input-output approach, in which a persistently exciting external signal is needed. For the proposed project, the error between quantized measurements and unquantized measurements is equivalent to the external excitation. Thus the same procedures as in Van den Hof and Schrama (1993) and Huang and Shah (1997) can be used.
- Bartee and McFarlane (1998) discussed the use of routine operating data without external excitation for closed-loop identification with simulation examples. The conclusion is that archived data has limited use in closed-loop identification. Therefore, there is a need for some kind of external excitation.
- Sun *et al.* (2001) proposed a new identification algorithm based on output over-sampling for linear discrete-time closed-loop system. This publication suggests that changing sampling rate for quantized signal processing may be a useful choice besides A/D converter of a different quantization interval.

2.2 Relay-based identification

In this section, the relay-based identification will be presented first, followed by the motivation. The emphasis will be placed on the development of relay-based identification schemes. The identifiability of the relay-based identification method, the noise issue and load disturbance impact, and the commercial products will also be briefly discussed.

2.2.1 Introduction

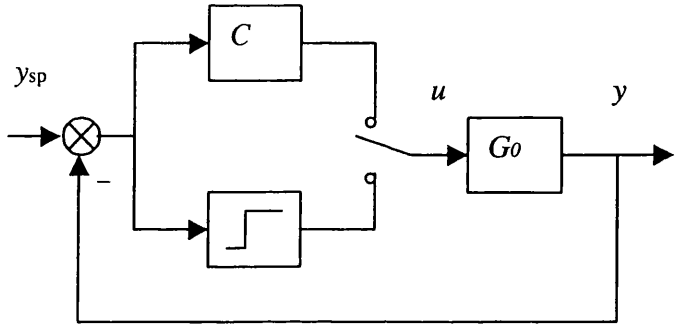


Figure 2-2 *Relay feedback system*

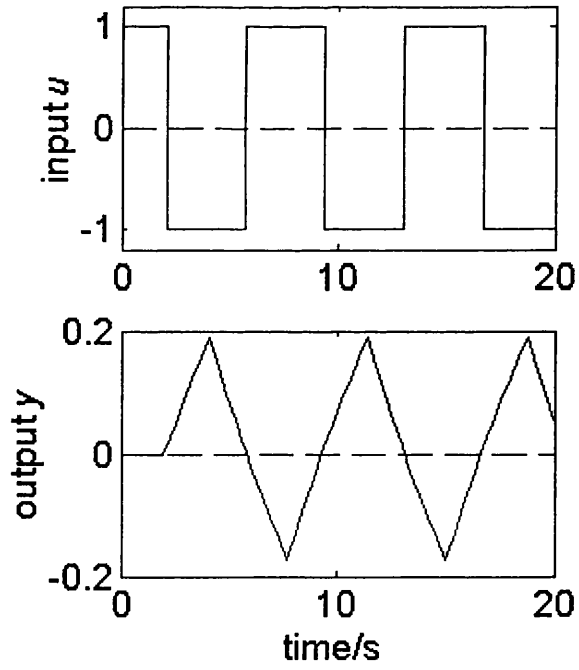


Figure 2-3 *Limit cycle from relay feedback, 'input u' means the process input; 'output y' means the process output.as shown in Figure 2-2.*

The relay-based identification method of Åström and Hägglund (1984a and 1984b) led to a new generation of process controllers having an auto-tuning facility. The idea was to introduce a nonlinear feedback of the ideal relay in order to generate a *limit cycle* oscillation (Figure 2-2). The system then starts to oscillate. The period and the amplitude of the oscillation are determined when steady-state oscillation is obtained (Figure 2-3). This gives the critical period and the critical gain (Åström and Hägglund, 1995).

2.2.2 Motivation

In frequency response methods, the process transfer function can be determined by measuring the steady-state responses to sinusoidal inputs. The result is a graphical model. When the frequency ω goes from 0 to ∞, the endpoint of the vector $G(j\omega)$ describes a curve in the complex plane, which is called the frequency curve or the Nyquist curve. The Nyquist curve gives a complete description of the system. Not every point in the Nyquist curve is so useful. For controller tuning, there are some parts that are of particular interest. For instance, the lowest frequency where the phase is -180° is called the critical frequency. The corresponding point on the Nyquist curve is called the *critical point* (Fig. 2-4). A difficulty with the traditional frequency response is that appropriate frequency of the input signal must be chosen *a priori*.

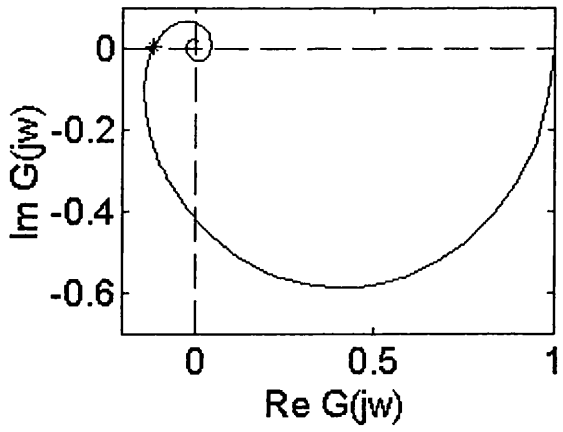


Figure 2-4 *Critical point (*) in the Nyquist curve of a first-order plus dead time process*

Ziegler and Nichols have provided a method for determining the critical point on the Nyquist curve experimentally. A system can be made unstable under proportional feedback by increasing the gain high enough. Thus there must be a gain that makes the system oscillate all the time. The gain is called the critical gain. This method is generally referred to as Ziegler-Nichols frequency response method, which can be interpreted as a method where one point (critical point) of the Nyquist curve is positioned. The advantage of the method is that the process itself is used to find the critical frequency. However, it is difficult to implement in practice since operating the process near instability is dangerous (Åström and Hägglund, 1995).

Auto-tuning (or automatic tuning) is a method where the controller is tuned automatically on demand from the user (Åström and Hägglund, 1995). Relay-based identification method was initiated in Åström and Hägglund (1984a) when they tried to find an easy and reliable auto-tuning facility for PID controllers. The idea was to replace the available controller with a nonlinear relay in order to generate limit cycle oscillations. The period and the amplitude of the oscillations can be determined when steady-state oscillation is obtained. This gives the critical period and the critical gain. These results can be used for automatic tuning with standard tuning tools, for example, Ziegler Nichols rules. Hang *et al.* (2002) gives a review on recent developments of relay identification and auto-tuning.

2.2.3 Framework of methods

In relay-based identification, the key points are

- The limit cycle generation mechanism, i.e. how to use the relay to generate limit cycle?
- The limit cycle analysis method (for example, the describing function analysis)
- The information obtained (for example, the point information in Nyquist plot)

Therefore, different methods can be classified according to the above criteria.

In more detail, the limit cycle generation mechanism can be

- Standard relay
- Amplitude biased relay in which the switch-on value and the switch-off value are not symmetric (Wang *et al.*, 1997a or Figure 3-12 and 3-13 of this thesis)
- Saturation relay in which the relay output is proportional to the input when the input to the relay is less than a specified value and the relay output is the relay amplitude when the input to the relay is greater than a specified value (Shen *et al.* 1996a)

- Set-point relay in which the relay is placed in the outer loop while the inner loop is the former loop (including both the controller and the process) (Luo *et al.*, 1998 and Schei, 1992).

The limit cycle analysis method can be one of the following

- Describing function analysis in which only the fundamental harmonic is considered and the higher-frequency harmonics can all be neglected (Atherton, 1975 or Section 3.3.1 of this thesis).
- Time-domain dynamic analysis solution in which the accurate transient behaviour (in time domain) of the process output corresponding to the relay output (as process input) is analyzed in detail (Wang *et al.*, 1997a).
- Fourier Transform (FFT) in which an exponential decay is introduced for both the input and the output at first, the process can then be obtained by dividing the Fourier transform of the output and the Fourier transform of the input (Wang *et al.*, 1997b).

The information obtained

- Frequency-domain one-point identification
- Frequency-domain two-point identification
- Frequency-domain multi-point identification

As an example to explain the above framework, the original relay-based identification method by Åström and Hägglund (1984a and 1984b) used a standard relay to generate limit cycle and the describing function analysis method to analyze the limit cycle. One-point information in frequency-domain was obtained for subsequent PID tuning.

The describing function analysis method will be described in detail in subsection 3.3.1 of this thesis. It will be briefly introduced here to provide some background material for further discussion in this chapter. In order to analyze the system represented in Figure 2-3, which includes an ideal relay — a nonlinear component, the concept of *describing function analysis* is used (see, for example, Atherton, 1975). By the describing function analysis method, the critical gain for the system $G(j\omega)$ represented in Figure 2-3 is

$$K_u = \frac{4d}{\pi a} \tag{2.7}$$

where d is the relay amplitude and a is the amplitude of the oscillation in the process output.

The critical frequency ω_u can be determined from the limit cycle by

$$\omega_u = \frac{2\pi}{P_u} \tag{2.8}$$

where P_u is the period of limit cycles.

2.2.4 Specific methods

This subsection concentrates on various methods to improve the original relay identification by Åström and Hägglund (1984a and 1984b). Different methods are arranged according to the information identified, for example, one point, two points, even multiple points in the Nyquist curve. Fig. 2-5 gives a graphical demonstration of the identified information by all the methods reviewed.

Frequency-domain one-point identification

Luyben (1987) used the relay-based method in conjunction with steady-state gain obtained independently, to find the best transfer function model for the processes. This is the first time researchers tried to identify transfer function model using relay-based method. A standard relay was used to generate the limit cycle and the describing function analysis method was employed to analyze the limit cycle. The limitation of Luyben (1987) is that the steady-state gain must be known ahead of the relay identification test.

Chang *et al.* (1992) modified Luyben (1987) to improve the accuracy of the model parameters. They proposed new accurate analytical expressions to calculate model parameters with the period of oscillation P_u and amplitude a observed from the limit cycle. The formula (2.7) and (2.8) derived from the describing function analysis method, which were used in Luyben (1987), are approximate because only the first harmonic is considered. Shen *et al.* (1996a) used saturation relay instead of the conventional ideal relay to improve the identification accuracy of the critical data.

Sung *et al.* (1995) used a six-step signal instead of the two-step signal (from an ideal relay) to reduce the harmonic terms when the signal is represented by the Fourier series. The proposed test signal looks similar as being quantized. More accurate critical data can be obtained from the test.

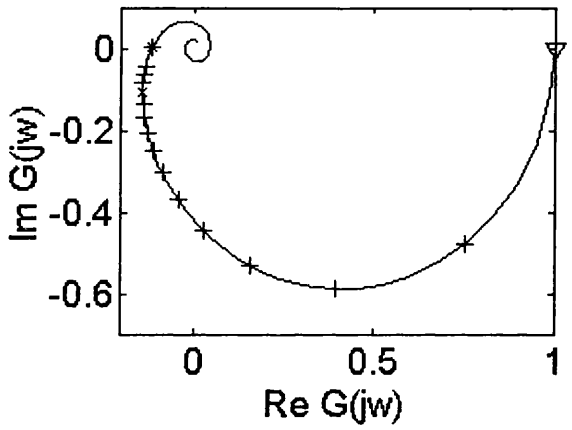


Figure 2-5 *Information identified by different relay-based identification methods (All frequency one-point identification methods identified the critical point* (*). *For frequency two-point identification methods, Li et al. (1991) identified the critical frequency* (*) *and one point in the third quadrant* (×), *Shen et al. (1996b) and Wang et al. (1997a) identified the critical frequency* (*) *and the zero frequency* (∇). *For frequency multiple-point identification, Wang et al. (1997b) and Bi et al. (1997) identified all the points in this figure including the points* (+).

Friman and Waller (1997) used two relays in the closed-loop, one acts on the error and the other acts on the integral of the error, to identify a point in the third quadrant of the Nyquist curve for better controller tuning. Besancon-Voda and Roux-Buisson (1997) proposed a different method for the identification of a point of prescribed phase lag on the process Nyquist curve by using an adjustable time delay instead of an adjustable relay hysteresis. A global optimization approach (such as: multi-pass Luus Jaakola method, modified Simplex method and a quasi-Newton method) for identifying a second-order-plus-dead-time (SOPDT) model from relay feedback is proposed by Kumar and Rangaiah (1999).

Frequency-domain two-point identification

Li *et al.* (1991) improved from Luyben (1987). They proposed a modified procedure that does not require knowledge of the steady-state gain. The method uses two tests. The first is a normal one with an ideal relay; the second is run with an ideal relay plus an additional known dead time so that the phase angle is shifted about 45° and a smaller critical frequency is obtained. From formula (2.7) and (2.8) - the result of the describing function analysis method, there are two equations from each test. From the four equations, a least-square method is then used to determine the unknown parameters: time constant(s) and the steady-state gain.

Based on the concept of dual-input describing function, Shen *et al.* (1996b) suggested only one test with a biased relay in order to obtain process information at two frequencies (the critical frequency and the zero frequency). The steady-state gain can be found from integrating system input-output response. Wang *et al.* (1997a) suggested an amplitude biased relay which needs only one test, but based on time-domain analysis (not describing function), to obtain the accurate (not approximate) first-order plus dead-time (FOPDT) model. The information identified is also two points of the Nyquist curve (the critical frequency and the zero frequency).

Frequency-domain multiple-point identification

Wang *et al.* (1997b) introduced an exponential decay into the standard relay output and the process output. With one relay test, multiple points of the frequency response can be obtained by Fourier transform. Bi *et al.* (1997) proposed another scheme for identifying multiple points of the frequency response. Wang *et al.*(1997b) still used the ideal relay, while a parasitic relay is superimposed to the standard relay in Bi *et al.* (1997) and the parasitic relay on-off period is twice as large as that of the standard relay.

Setpoint relay

Luo *et al.* (1998) introduced the set-point relay to extract relevant information about the process dynamics. There is no need to switch off the existing controller, which is a benefit because this means less disruption to the process. A similar scheme based on setpoint relay is Schei (1992). The model identification and control design are performed via on-line iteration. A shortcoming of Schei (1992) is that the on-line experiment time is very long due to iteration Luo *et al.* (1998).

Phase locked loop

Crowe and Johnson (2000) proposed a phase-locked loop framework as an alternative to the relay method for extracting frequency response. The main components are:

- A feedback structure with a phase or gain reference as an input comparator.
- A digital oscillator to generate a process sinusoidal extraction path and a sinusoidal reference path.
- A digital signal processing unit to extract the measured system phase or gain for use as another input to the comparator.
- A digital integrator unit to guarantee the identifier unit converge to the given phase or gain reference

Automated accurate frequency response can be achieved with this method. The drawback of the method is the long estimation times needed for some tests.

2.2.5 Identifiability, noise issue and load disturbance

Identifiability

The key to relay-based identification is generation of the stable limit cycles. Åström and Hägglund (1984a) showed that stable periodic solutions will not be obtained for all systems. The necessary condition for local stability of limit cycle in Åström and Hägglund (1984b) give some guidance when there will be limit cycle oscillations and when the limit cycle oscillations are stable, but the general conditions are still unknown.

Noise issue

For practical use, the measurement noise influence is inevitable in relay-based identification because the process variables obtained from the sensors generally include measurement noise. To avoid the measurement noise influence, Åström and Hägglund (1984a) suggested two measures. A hysteresis in the relay is a simple way to reduce the influence of measurement noise. The width of hysteresis should be bigger than the noise band and is usually chosen as 2 times larger than the noise band (Åström and Hägglund, 1995). Filtering is another way. The measurement noise is usually of high frequency, while the process frequency response of interest is usually in low frequency. A Butterworth lowpass filter was used in Wang *et al.* (1997a) and Wang *et al.* (1997b) respectively to prevent noise influence.

Load disturbance

> Load disturbances are the disturbances that enter the control loop somewhere in the process and drive the system away from its desired operating point.
>
> (Åström and Hägglund, 1995, 54)

Load disturbances can be modeled as step function added to the control signal at the process input because they are typically low-frequency disturbances, for example, it means *d* is a step in Figure 2-1. Load disturbances during the relay experiment introduce errors in the estimates of the critical gain and critical period. Hang *et al.* (1993) proposed the use of an automatically biased (displaced) relay feedback test to overcome static input-side load disturbance. This method needs prior appropriate information on the static gain of the process. To overcome this problem, Park *et al.* (1997) suggested a new relay feedback method, which shifts the relay output automatically to guarantee the symmetric process output. Leonard and Oubrahim (1998) improved Hang *et al.* (1993) by proposing a two-step relay test. They consist of centering the relay directly on the average output of a PID controller already in the loop for symmetric oscillation cycles in the presence of the input-side load disturbance. Shen *et al.* (1996c) used an output-biased relay to identify the process in the presence of an output-side load disturbance.

2.2.6 Products

From Åström *et al.* (1993) and Åström and Hägglund (1995), the Afa Laval Automation ECA400 controller that is commercially available is based on relay identification. Wang *et al.* (1999) reported the laboratory implementation of Wang et al. (1997b) using a personal computer. Report on commercial product from Wang et al. (1997b) has not been found yet.

2.2.7 Conclusions

From the discussions in this section, the following conclusions can be reached:

- The relay-based identification indicates the trend of identifying information in the frequency-domain from one-point to two-point and multi-point.
- Wang *et al.* (1997b) is a mature method. It can obtain multiple-point information in the Nyquist plot with an ideal relay to generate a limit cycle. An exponential is then applied to the process input and output respectively. Fourier transform is subsequently used to obtain the process response in the frequency domain.

- Luo *et al.* (1998) proposed the set-point relay, in which during the identification experiment there is no need to switch off the existing controller. The combination of Wang *et al.* (1997b) and Luo *et al.* (1998) provide some possibility of a new relay identification scheme using quantization signal.
- The measurement noise is common in industrial practice. It is necessary to see noise influence for a specific relay-based identification method.

2.3 Other relevant CLID methods

The following publications are important for this study, but cannot be categorized into the above branches.

2.3.1 CLID based on quantization

Welsh and Goodwin (1999) proposed a novel autotuning method based on quantization. In essence, this is the use of a quantizer at the controller output, instead of a relay, for identification. The main advantage over relay-based identification is that the controller is kept in the control loop during the identification phase since compared to other methods where the replacement of the PID controller by the relay disrupts the operation of the control loop. Goodwin and Welsh (1999) extended Welsh and Goodwin (1999) into multivariable autotuning. These two publications are helpful in defining the targets of this research (closed loop identification with a quantizer).

2.3.2 CLID based on multi-harmonic perturbation exciting

Fox and Godfrey (1998) proposed a nonparametric closed-loop identification method with multi-harmonic perturbation exciting (reference signal). The exciting signal must be carefully designed to excite the system response around the system critical frequency. This kind of perturbation is between the classical step input and the persistently exciting signal used in general identification. This is useful in quantization signal processing since in relay-based identification describing function analysis disintegrate the signal as combination as multi-harmonic signals.

2.3.3 Conclusion

- Welsh and Goodwin (1999) is relevant to the content of this thesis since a quantizer is used for closed-loop identification. It is important to avoid the disadvantage of replacing the controller with a relay.

2.4 Closed-loop performance assessment

In this section, the relevance of closed-loop performance assessment (CLPA) to closed-loop identification is first stated. Then, the development of closed-loop performance assessment will be reviewed.

2.4.1 Why closed-loop performance assessment is relevant to this thesis

The result of CLID is a dynamic model of the controlled process. The model is subsequently used for control design in most cases. The accuracy of the model is important. However, more important is the performance of the whole loop. Therefore, CLPA is a quite relevant topic with CLID. It is useful to review CLPA in the context of CLID. Moreover, the CLPA theory will be used for the new theory of CLID based on the quantization in Chapter 5 and Chapter 6.

2.4.2 Traditional performance assessment

According to Stephanopoulos (1984), when the input to the system is a set-point step disturbance or load upset variable, the performance criteria were clarified as steady-state performance criteria and dynamic response performance criteria. Furthermore, Dynamic response Performance Criteria were divided into simple performance criteria (such as overshoot, rise time, settling time, decay ratio) and time-integral performance criteria (e.g. integral of the square error (ISE), integral of the absolute value of the error (IAE), integral of the time weighted absolute error (ITAE)). The simple performance criteria uses only isolated characteristics of the dynamic response. The time-integral performance criteria are based on the entire response of the process. That is

- Steady-state performance criteria
- Dynamic response performance criteria
    - Simple performance criteria
    - Time-integral performance criteria

2.4.3 Modern performance assessment overview

The performance of an existing control loop is often measured against some kind of benchmark. There are many different measures of control performance. Harris *et al.* (1999)

gave an overview on performance monitoring and assessment techniques. Qin (1998) reviewed the current status of control performance monitoring using Minimum Variance Control (MVC). Tradeoffs between performance and robustness, stochastic and deterministic performance objectives are also discussed. For instance, it is common practice to design an internal model control (IMC) filter to make the controller robust. The larger the time constant for the filter, the more robust the controller, the IMC-achievable variance will dramatically increase. This is the tradeoff between performance and robustness.

2.4.4 Closed-loop performance assessment index and its calculation

Harris (1989) developed an efficient technique for control performance assessment using only routine closed-loop operating data. Minimum variance control is used as the benchmark standard against which to assess current control loop performance. Many authors regarded this paper as a milestone. The advantage of Harris (1989) is that it is not necessary to perturb the process to obtain the necessary information for analysis. Desborough and Harris (1992) implemented the idea described above. A normalized performance index $\eta(b)$ (b is how far ahead to do the forward prediction) was introduced to characterize the performance of feedback control schemes. The normalized performance index $\eta(b)$ can be estimated by linear regression methods.

According to Harris (1989) and Desborough and Harris (1992), in Figure 2-6, the controller error e can be expressed as

$$e = \hat{e} + r \tag{2.9}$$

where $\hat{e}$ is forward prediction of the controller error and r is the residual.

The closed-loop performance assessment (CLPA) index was defined as:

$$\eta(b) = 1 - \frac{\sigma_{MV}^2}{\sigma_e^2} \tag{2.10}$$

where b is the forward prediction (b is time delay if it is known, otherwise b is an estimate of the time delay), σ_e^2 is the variance of the controller error and σ_{MV}^2 is the variance of (zero mean) residuals (i.e. the minimum variance achievable).

The qualitative relation is that the smaller the predictable component $\hat{e}$, the better the controller, where $\eta(b)$ tends to zero.

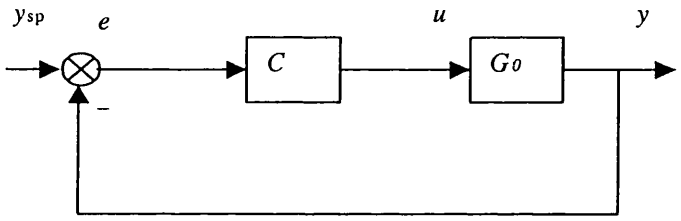


Figure 2-6 *Closed-loop system*

2.4.5 Modern performance assessment development

Development of performance index proposed by Harris

Stanfelj *et al.* (1993) extended this technique to feedforward/feedback loops. They developed a cross-correlation test to determine when a feedforward control loop is performing optimally in the mean square sense. At the same time, they developed a cross-correlation analysis technique to identify whether the feedforward or the feedback component of the control loop was responsible for sub-optimal performance. Both of the tests use routine operating data. Thornhill *et al.* (1999) discussed the application of the above method to a refinery. Practical experiences with the use of the techniques were reported.

Lynch and Dumont (1996) have also reported the use of minimum variance control as a benchmark, but they used a modeling method known as Laguerre networks in modeling the controller output, instead of an ARMA model by Harris (1989), to determine the minimum variance of output error.

The normalized performance index may be estimated online using recursive least squares, enabling the use of control charts to monitor performance changes online (Desborough and

Harris, 1992). Jofriet *et al.* (1996) fulfilled on-line performance assessment for an individual control loop using an expert system. The expert system implemented in Gensym's G2, still uses the normalized variance-based performance index proposed by Harris in 1992.

Even though these 'minimum variance' methods may provide useful information about achievable limits without perturbing the process externally, they have several shortcomings as tools for evaluating deteriorating control (Bezergianni and Georgakis, 2000). First, achieving this theoretical limit may require a controller with high bandwidth or excessive control action and many loops do not operate at minimum variance. Second, a shift in the ratio of actual variance to minimum variance may be due to either changes in the controller, changes in the plant, or changes in the disturbance spectrum. Whereas changes in the controller or the plant may merit re-tuning, changes in the disturbance spectrum may not.

Performance assessment and system identification

Closed loop performance assessment techniques as discussed above do not require an explicit process model when minimum variance control is used as a benchmark. However, Huang (1997) pointed out that a prerequisite for closed loop performance assessment at a higher level is generally a model of the process. Performance assessment with a benchmark other than minimum variance control (MVC) usually requires an identification effort. In fact, apart from variance-based criteria, other criteria for assessing and diagnosing controller performance have been studied, for example, LQG benchmark, user-defined benchmark.

Bezergianni and Georgakis (2000) defined a new performance index called relative variance index (RVI) after they analyzed the limitations of the minimum variance control (MVC) methodology. The proposed new index compared the current output variance under closed loop conditions with the variance under the 'best' (i.e. MVC) and under the 'worst' (i.e. open loop) possible control action. A variable regression estimation (VRE) technique was utilized for the evaluation of the number of process unit delays. Structured target factor analysis (STFA) and subspace identification were used for the identification of the controller model and process model from operating data. The RVI estimation from closed loop operating data is more accurate when the identified models are closer to the actual system models. The method needs at least a set-point change.

Kendra and Cinar (1997) proposed a system identification based method for assessing the performance of closed-loop system, utilizing measures which coincide naturally with classical

and modern frequency domain design specifications, such as the bandwidth, peak magnitude of the sensitivity and complementary sensitivity function. Estimates of these transfer functions can be obtained by exciting the reference input with a zero mean PRBS (Pseudo Random Binary Signal), observing the process output and error response, and developing a closed-loop model. Performance assessment is based on the comparison between the observed frequency response characteristics and the design specifications.

Until now, there is no report on closed loop performance assessment with closed loop identification using operating data without external excitation.

2.4.6 Performance assessment for PID controller

Some other progress in closed loop performance assessment includes researchers' efforts specifically for PID controllers. This does not mean the performance measures discussed above are not suitable for PID controller. In fact, they are suitable. But the performance measures discussed here are only suitable for control loops with PID controllers.

Åström (1991) studied qualitative and quantitative assessment of the performance that can be achieved with a simple feedback loop from the viewpoint of expert control. The qualitative measures (such as: stable/unstable, monotone/oscillatory) and quantitative measures (such as: bandwidth, peak error, maximum time, integral gain) can be assessed from a step test or from an experiment with relay feedback. A more accurate assessment of the control performance can be made if the transfer function of the process is known.

For simple feedback loops with PID controllers that are tuned using Ziegler-Nichols formula, if the step response is monotone or essentially monotone, achieved performance can be characterized with dimensionless numbers such as normalized deadtime, normalized process gain, peak load error and normalized closed-loop rise time (Åström et al. 1992). The last two quantities can be assessed online. Ko and Edgar (1998) proposed the specific procedures in estimating the achievable PI control performance when the process was affected by stochastic load disturbances.

2.4.7 Conclusions

- The measure defined by Harris (1989) is viewed as a milestone for control performance assessment. Minimum variance control is used as the benchmark. The advantage of Harris (1989) is that it is not necessary to perturb the process to obtain the necessary information for analysis.
- Desborough and Harris (1992) used routine closed-loop process data to estimate the normalized performance index with least square method.
- Ko and Edgar (1998) proposed the specific procedures in estimating the achievable PI control performance when the process was affected by stochastic load disturbances. This is the application of the performance index proposed by Harris (1989) in PI control. This publication will be useful since PID controller is selected for the proposed project and there is a need to demonstrate better performance achieved than before by experimentation.

2.5 Conclusions

In this chapter, the closed-loop identification with external excitation method and the relay-based identification method have been reviewed in detail. Other CLID methods such as CLID based on quantization have been involved. The relevant topic of closed-loop performance assessment has also been discussed.

- Closed-loop identifiability (Gustavsson *et al.*, 1977; MacGregor and Fogal, 1995 and Bartee and McFarlane, 1998) needs to be studied theoretically and by simulation example since the insight into closed-loop identifiabilty is important in further work on this thesis.
- With regard to closed-loop identification with external excitation method, the two-stage method (Van den Hof and Schrama, 1993) and two-step method (Huang and Shah, 1997) are practical methods that change the closed-loop identification problem into two open-loop identification problems.
- For relay-based identification, the trend of identifying frequency-domain information from one-point (Åström and Hägglund, 1984a and 1984b) to two-point (Li *et al.*, 1991; Wang *et al.*, 1997a) to multi-point (Wang *et al.*, 1997b) has been pointed out.
- Closed-loop identification based on quantization (Goodwin and Welsh, 1999 and Welsh and Goodwin, 1999) is relevant to this thesis.
- Control performance assessment (CLPA) defined by Harris (1989) is a milestone. Desborough and Harris (1992) used routine closed-loop process data to estimate the normalized CLPA index with least square method.

CHAPTER 3

CASE STUDIES

This chapter describes work on a series of case studies. The motivation for conducting these studies was to maximize personal understanding of closed loop identification and prepare closed loop identification schemes for further use.

The studies include:

- Closed loop identifiability
  - MacGregor and Fogal (1995), Closed-loop identification: the role of noise model and prefilters
- Closed loop identification with external excitation
  - Van den Hof and Scharama (1993), An indirect method for transfer function estimation from closed loop data
  - Huang and Shah (1997), Closed-loop identification: a two step approach
- Relay-based identification
  - Li, *et al.* (1991), An improved autotune identification method
  - Wang *et al.* (1997a), Low-order modeling from relay feedback
  - Wang *et al.* (1997b), Process frequency response estimation from relay feedback

This chapter is laid out as follows. Closed loop identifiability and noise model influence are first examined. Then two specific methods of closed loop identification with external excitation and three specific methods of relay-based identification are introduced and duplicated. The two main categories are subsequently compared. This is the attempt to make a connection of previously unrelated facts. The conclusion is given at the end.

3.1 Closed-loop identifiability and influence of the noise model

In this section, the theory on closed-loop identifiability will be presented first. Simulation examples will follow. The conclusion will be given based on the results and discussions.

3.1.1 Theory

In this subsection, the theory on closed-loop identification using nonparametric methods, closed-loop identification using parametric methods and the influence of noise model will first be discussed. Then the asymptotic theory for closed-loop identification will be derived to explain the conclusions above.

As shown in subsection 1.2.4, the SISO closed-loop system (adapted from MacGregor and Fogal, 1995) is presented in Figure 3-1, which is the same as Figure 1-2. It provides a general structure for CLID with external excitation. External excitation is a dither signal injected into the original closed-loop system to excite the process for system identification (*d* is Figure 3-1).

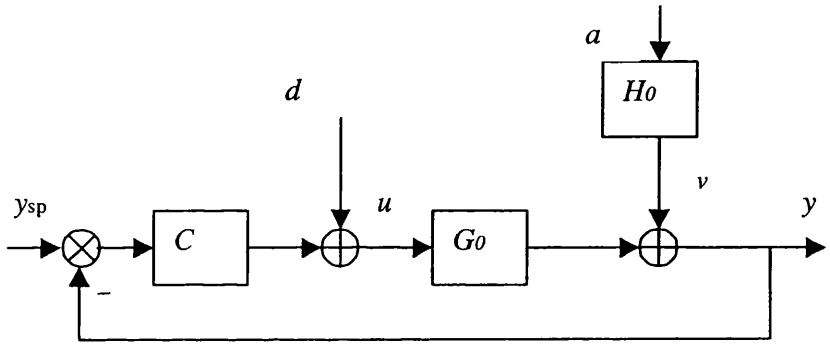


Figure 3-1 *Closed-loop system*

The true system is assumed to be

$$y(t) = G_0(q)u(t) + H_0(q)a(t) \qquad (3.1)$$

where $G_0(q)$ represents the true process, and the disturbance $v(t) = H_0(q)a(t)$ gives the effect of all unmeasured process disturbances on the measured output *y(t)*, *q* is the forward shift operator.

$$u(t) = C(q)(y_{sp}(t) - y(t)) + d(t) \qquad (3.2)$$

The feedback loop is given in (3.2) in which *C(q)* is the feedback controller. The set-point $y_{sp}(t)$ and the 'dither' signal *d(t)* are the possible input signals that will be injected for

identification. Only $y_{sp}(t) = 0$ is considered in this thesis, that is to say, only use *d(t)* as an input signal because the system considered is assumed to be a linear system, the output of the system is the superposition of the response to $y_{sp}(t)$ and response to *d(t)* if $y_{sp}(t)$ and *d(t)* are applied simultaneously. Thus the feedback equation (3.2) becomes

$$u(t) = -C(q)y(t) + d(t) \tag{3.3}$$

Throughout this thesis, assume that the system model is given by

$$y(t) = G(q,\rho)u(t) + H(q,\eta)a(t) \tag{3.4}$$

where ρ and η are vectors of parameters and $a(t)$ represents the innovation sequence.

For the system described above, the challenge is to identify the process while the process is under feedback control and is excited by an external excitation *d(t)*.

Closed-loop identifiability using non-parametric methods

As mentioned earlier, *system identifiability* refers to the identifiability of both the process $G_0(q)$ and the disturbance/noise model $H_0(q)$. However, throughout this thesis, only the identifiability of the process $G_0(q)$ is concerned because only the true process is of interest.

For the closed-loop system described by equation (3.1) and equation (3.3), using non-parametric methods on closed-loop data (for example, the spectral analysis method), the frequency response function can be given by (MacGregor and Fogal, 1995)

$$G(\omega) = \frac{\phi_{uy}(\omega)}{\phi_u(\omega)} = \frac{G_0(e^{j\omega})\phi_d(\omega) - C(e^{j\omega})\left|H_0(e^{j\omega})\right|^2 \sigma_a{}^2}{\phi_d(\omega) + C(e^{j\omega})^2\left|H_0(e^{j\omega})\right|^2 \sigma_a{}^2} \tag{3.5}$$

where ϕ_{uy} and ϕ_u are the cross-spectrum *u(t)* and *y(t)* and the spectrum of *u(t)* respectively, ϕ_d is the spectrum of *d(t)* and $\sigma_a{}^2$ is the variance of white noise *a(t)*. Several important results can be derived from (3.5).

- If no external excitation is injected (i.e. $\phi_d = 0$), equation (3.5) becomes

$$G(\omega) = -\frac{1}{C(e^{j\omega})} \tag{3.6}$$

This expression shows that when the process is operating under pure feedback (i.e. no external excitation), the non-parametric method identifies only the inverse of the feedback controller, and gives no information on the real process. Therefore, it is impossible to do CLID without external excitation using nonparametric methods.

- If an external dither signal is injected (i.e. $\phi_d \neq 0$), according to (3.5), the estimates give a weighted average of the real process frequency response and the frequency response of the inverse of the feedback controller. The weighting factor depends upon the signal to noise ratio (*S/N*), i.e. $\sigma_d^2 / \sigma_v^2 = \sigma_d^2 / \left|H_0(e^{j\omega})\right|^2 \sigma_a^2$.

Closed-loop identifiability using parametric methods

Under pure feedback:

When parametric methods are used to deal with data collected from the system operating under pure feedback, MacGregor and Fogal (1995) indicated that parametric models can only be identified if fairly restrictive necessary and sufficient conditions on the real process and controller orders are satisfied under pure feedback conditions.

For a real process $G_0(q)$ represented by an ARMAX model:

$$\mathrm{A}(q^{-1})\, y(t) = q^{-k}\mathrm{B}(q^{-1})\, u(t) + \mathrm{C}(q^{-1})\, a(t) \tag{3.7}$$

With feedback controller $\mathrm{C}(q^{-1})$:

$$\mathrm{F}(q^{-1})u(t) = \mathrm{G}(q^{-1})y(t) \tag{3.8}$$

The necessary and sufficient condition for asymptotic identifiability (see MacGregor and Fogal, 1995) is that:

$$\max [n_f - n_b, k + n_g - n_a - 1] \geq n_p \qquad (3.9)$$

where n_a, n_b and k are the orders of the polynomials in the real process $G_0(q^{-1})$, n_f and n_g are the orders of the controller polynomials $C(q^{-1})$, n_p is the order of the common factor in the polynomials $C(q^{-1})$ and $A(q^{-1})F(q^{-1}) - q^{-k}B(q^{-1})G(q^{-1})$.

This condition could be satisfied when there are sufficiently large delay (k) in the process and/or when the controller used is of sufficiently high order. Otherwise, if this condition is not satisfied, then no model of any structure that adequately describes the real process can be identified under pure feedback conditions.

With external excitation

As discussed in Subsection 2.1.5, something must be done to break the dependency between the input signal and the process disturbance for closed-loop identification (MacGregor and Fogal, 1995; Bartee and McFarlane, 1998).

(1) By injecting an independent, persistently exciting signal into the feedback loop.

(2) By switching between two or more feedback controllers.

Either of these will guarantee that necessary and sufficient conditions for identifiability be satisfied.

Noise model influence in closed-loop identification

For open-loop identification, Ljung (1987) concluded that even if the noise model $H_0(q)$ (as expressed in equation 3.1 and in Figure 3-1) is incorrect, as long as it is stationary, the true process can be asymptotically identified using a sufficiently rich process model (i.e. the assumed model structure can contain the true process). However, this is not true for closed-loop identification.

MacGregor and Fogal (1995) studied this theoretically and by simulations. The main points are: In closed-loop identification (using *direct* method), the correct process model can only be identified if one also identifies the correct noise model. If a pre-filter is used to filter the input-output data, then the pre-filter should be exactly the inverse of the true process noise model so that an accurate process model can be obtained.

Asymptotic theory for closed-loop identification

The system in Figure 3-1 is considered and the system model given in (3.4) is assumed, the model residuals or prediction errors can be derived as:

$$e(t,\rho,\eta) = H^{-1}(q,\eta)[y(t) - G(q,\rho)u(t)] \tag{3.10}$$

From the system described by (3.1) and (3.3), the sensitivity function of the real closed-loop system can be obtained.

$$S_0(q) = [1 + G_0(q)C(q)]^{-1} \tag{3.11}$$

Using (3.11), (3.1) and (3.3), the system can be expressed as the following:

$$y(t) = G_0(q)S_0(q)d(t) + S_0(q)v(t) \tag{3.12}$$

$$u(t) = S_0(q)[d(t) - C(q)v(t)] \tag{3.13}$$

Therefore, the spectrum of the prediction errors is given by (MacGregor and Fogal, 1995):

$$\phi_e(\omega,\rho,\eta) = \{\left|G_0(e^{j\omega}) - G(e^{j\omega},\rho)\right|^2 \phi_d(\omega) + \left|1 + G(e^{j\omega},\rho)C(e^{j\omega})\right|^2 \phi_v(\omega)\} \frac{\left|S_0(e^{j\omega})\right|^2}{\left|H(e^{j\omega},\eta)\right|^2} \tag{3.14}$$

In (3.14), the sensitivity function of the estimated closed-loop system can be written as:

$$S(e^{j\omega},\rho) = [1 + G(e^{j\omega},\rho)C(e^{j\omega})]^{-1} \tag{3.15}$$

If the criterion of minimizing the sum of squares of the prediction errors is chosen for the estimation of parameters ρ and η, then the objective function is given by

$$V(\rho,\eta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_e(\omega,\rho,\eta)d\omega \tag{3.16}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|G_0(e^{j\omega}) - G(e^{j\omega},\rho)\right|^2 \cdot \frac{\left|S_0(e^{j\omega})\right|^2 \phi_d(\omega)}{\left|H(e^{j\omega},\eta)\right|^2} d\omega$$

$$+ \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|S_0(e^{j\omega})\right|^2}{\left|S(e^{j\omega},\rho)\right|^2} \cdot \frac{\left|H_0(e^{j\omega})\right|^2 \sigma_a^{\,2}}{\left|H(e^{j\omega},\eta)\right|^2} d\omega \qquad (3.17)$$

From (3.17), the asymptotic prediction error variance consists of two terms:

- The first term —the bias term: a term relevant to the bias in the process model $\left|G_0(e^{j\omega}) - G(e^{j\omega},\rho)\right|^2$ weighted by a frequency term dependent on the ratio of the spectrum of the external signal to that of the noise model $\phi_d(\omega)/\left|H(e^{j\omega},\eta)\right|^2$.
- The second term —the sensitivity ratio term: a term relevant to the ratio of the real and the estimated closed-loop sensitivity functions, weighted by the ratio of the real noise spectrum to the spectrum of the estimated noise.

The expression (3.17) is one of the most important expressions in closed-loop identification. It can be used to illustrate the topics discussed above: Closed-loop identifiability using non-parametric methods and closed-loop identifiability using parametric methods. It will also be used to predict the noise model influence in closed-loop identification. Several phenomena discussed earlier can be explained:

- *Nonparametric methods under pure feedback:*

  Under pure feedback (i.e. ϕ_d=0), the bias term (i.e. the first term) in (3.17) is zero no matter what the bias $\left|G_0(e^{j\omega}) - G(e^{j\omega},\rho)\right|$ is. Then the identification depends upon the minimization of the second term. The second term in (3.17) is minimized to be zero only when $1 + G(e^{j\omega},\rho)C(e^{j\omega}) = 0$. This will result in the same answer as (3.6). That is why under pure feedback, a non-parametric method can only find the inverse of the feedback controller instead of the real process.

- *Parametric methods under pure feedback:*

  Under pure feedback (i.e. ϕ_d=0), the bias term in (3.17) is still zero no matter what the bias $\left|G_0(e^{j\omega}) - G(e^{j\omega},\rho)\right|$ is. For parametric methods, the minimum of the second

term is σ_a^2 when the necessary and sufficient condition for asymptotical identifiability is satisfied. This minimum occurs on two occasions. One is the correct solution $G(e^{j\omega},\rho)=G_0(e^{j\omega})$ and $H(e^{j\omega},\eta)=H_0(e^{j\omega})$. Another one occurs for any other combination of $G(e^{j\omega},\rho)$ and $H(e^{j\omega},\eta)$ which make the term in the integral equal to unity. This is the reason why under pure feedback a model for the process can be identified, but the identified model cannot be validated.

- *Identification with external excitation:*

  When an external input is injected (i.e. $\phi_d \neq 0$), both terms in (3.17) will not become zero as before. If the noise model and the process model are rich enough (i.e. the assumed model structure can contain the true process, then the minimum of the first term is zero and the minimum of the second term is σ_a^2. They both occur only when $G(e^{j\omega},\rho)=G_0(e^{j\omega})$ and $H(e^{j\omega},\eta)=H_0(e^{j\omega})$.

- *Influence of the noise model*

  Otherwise, if the process noise model is not rich enough to capture the real process noise characteristics (i.e. $H(e^{j\omega},\eta)\neq H_0(e^{j\omega})$), then the second term cannot achieve σ_a^2 when the process estimate is correct (i.e. $G(e^{j\omega},\rho)=G_0(e^{j\omega})$). Thus a bias of the process estimate (i.e. $G(e^{j\omega},\rho)-G_0(e^{j\omega})$) must exist. Furthermore, the bias of the process estimate (i.e. $G(e^{j\omega},\rho)-G_0(e^{j\omega})$) has a direct relation with the signal to noise ratio. If the signal to noise ratio ($S/N = \sigma_d^2/\sigma_v^2$) is large, the first term in (3.17) will be dominant in the effort to minimize the sum of squares of the prediction errors. The identification would be like an open-loop identification problem. The bias will be relatively small. That is to say when the signal to noise ratio (*S/N*) is large enough, the process model bias will disappear asymptotically. On the other hand, if the signal-to-noise ratio (*S/N*) is small, the bias will be relatively large.

3.1.2 Simulation examples and cases

To demonstrate the theoretical analyses discussed above, a second order ARMAX model is chosen. The transfer function given by Huang and Shah (1997) is used here.

$$y(t) = \frac{0.3403 + 0.2417q^{-1}}{1 - 0.7859q^{-1} + 0.3679q^{-2}} u(t-1) + \frac{1 - 0.8q^{-1} + 0.12q^{-2}}{1 - 0.7859q^{-1} + 0.3679q^{-2}} a(t) \quad (3.18)$$

A unit feedback control law is implemented in this simulation, i.e. $C = 1$ in Figure 3-1. The white noise $a(t)$ and the white noise dither signal $d(t)$ are independent with each other. The number of data points used in the simulation is 5,000. The following cases were designed.

Case 1

The aim of this case is to demonstrate the theory *closed-loop identifiability using non-parametric method.* The Correlation analysis method, one of the typical non-parametric identification methods, was used in this case to find the step response. Different dither signals were injected respectively, ($\sigma_d^2 = 0$, $\sigma_d^2 = 1$ and $\sigma_d^2 = 9$) while keeping $\sigma_a^2 = 2.25$. This led to different signal to noise ratios ($S/N = \sigma_d^2 / \sigma_v^2$). The identified step responses under different signal to noise ratios will be compared to see whether the conclusion is in agreement with the theory discussed.

Case 2

This case is aimed at demonstrating *closed-loop identifiability using parametric method.* Prediction error method with a Box-Jenkins model (see Subsection 1.1.2), one of the typical parametric identification methods, was used. Different dither signals ($\sigma_d^2 = 0$, $\sigma_d^2 = 1$ and $\sigma_d^2 = 9$) while keeping the noise variance $\sigma_a^2 = 2.25$ led to different signal to noise ratios. The identified step responses and frequency responses (Nyquist plot) were compared to see whether it agrees with the theoretical part.

Case 3

The aim of this case is to validate the theory on the influence of the noise model in closed-loop identification using direct approach (see Subsection 2.1.3). Here, prediction error method with a Box-Jenkins model was used. If a first-order noise model is presumed (i.e. the noise model is different from the true noise model and the noise model is not rich enough to catch the true noise characteristics), what will the result be and what is the relationship between accuracy and the signal to noise ratio? In this case, two different dither signals were used respectively ($\sigma_d^2 = 1$, $\sigma_d^2 = 9$) while keeping the noise variance $\sigma_a^2 = 2.25$.

3.1.3 Results and discussions

Case 1

This case is to validate the theory *closed-loop identifiability using non-parametric method.* Under different conditions: pure feedback, small signal to noise ratio and large signal to noise ratio (as defined in Table 3.1), the estimates (step responses) using correlation analysis method are illustrated in Figure 3-2. Here only the step responses are illustrated because the output of the correlation analysis method is impulse response instead of frequency response.

Test Condition	Dither signal σ_d^2	Noise σ_a^2	$S/N = \sigma_d^2 / \sigma_v^2$
Pure feedback	0	2.25	0.000
Small signal to noise ratio	1	2.25	0.410
Large signal to noise ratio	9	2.25	3.685

Table 3.1 *Different signal to noise ratio conditions*

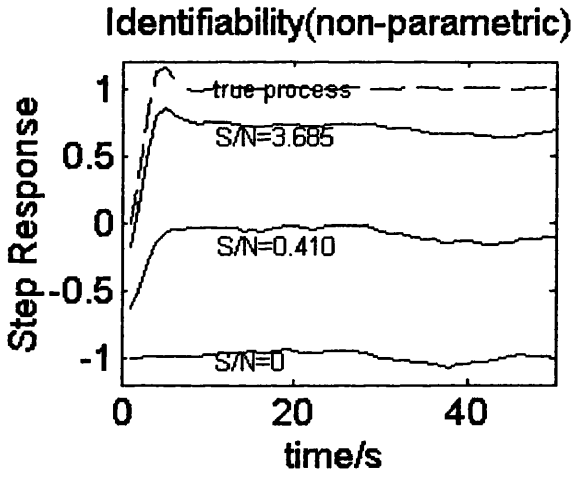


Figure 3-2 *Closed-loop identifiability using non-parametric method*

In Figure 3-2, the lower line was tested under pure feedback (i.e. $S/N = 0$), the model identified only reflected the inverse of the unit feedback controller. The upper line corresponds to large signal to noise ratio (i.e. $S/N = 3.685$). For the middle line, the small signal to noise ratio (i.e. $S/N = 0.410$) is between pure feedback and large signal to noise ratio, the estimated model is also between these two occasions. The result will be explained in the next section.

Closed-loop identifiability using non-parametric methods

The results in Figure 3-2 indicate the correctness of the theoretical analysis in section 3.1.1. The theory said that when the process is operating under pure feedback (i.e. no external excitation), the non-parametric method identifies only the inverse of the feedback controller, and gives no information on the real process. The results in Figure 3-2 show the same thing. From Figure 3-2, if no external excitation is injected (i.e. $S/N = 0$), the identified model only reflected the inverse of the feedback controller. The theory also said that when there is an external injection, the estimate is a weighted average of the real process and the inverse of the feedback controller. In Figure 3-2, the model identified with large signal to noise ratio (i.e. $S/N = 3.685$) reflected the real process basically. The model with moderate signal to noise ratio (i.e. $S/N = 0.410$) reflected the true process more poorly than the large signal to noise occasion.

Case 2

The purpose of this case is to demonstrate the theory of *closed-loop identifiability using parametric method.* The results shown in Figure 3-3 were identified using parametric methods under different situations (pure feedback, small signal to noise ratio and large signal to noise ratio as defined in Table 3.1).

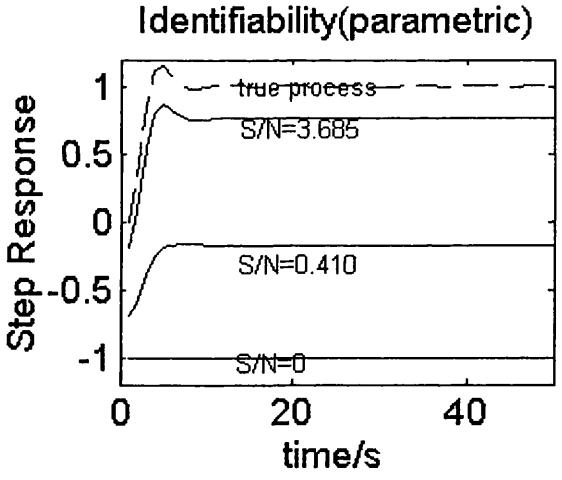


Figure 3-3 *Closed-loop identifiability using parametric method*

In Figure 3-3, the lower line was tested under pure feedback (i.e. $S/N = 0$), the estimate is a straight line, the identified model only reflected the inverse of the feedback controller. Discussion will show that the reason is the necessary and sufficient condition for asymptotic identifiability (3.9) cannot be satisfied under pure feedback. The upper line corresponds to large

signal to noise ratio (i.e. *S/N* = 3.685), it roughly reflected the real process. For the middle line, the signal to noise ratio (i.e. *S/N* = 0.410) is between pure feedback and large signal to noise ratio.

Compared with Figure 3-2, the step response values in Figure 3-3 are similar, but the step response estimates are smoother. That is because Figure 3-2 is obtained form the estimated results with correlation analysis method, one of the nonparametric methods and Figure 3-3 is obtained form the estimated results with prediction error method using ARX model, one of the parametric methods.

Closed-loop identifiability using parametric methods

For the process (described in equation 3.18) in subsection 3.1.2, there are two autoregressive terms, two moving average terms and delay of 1 (n_a=2, n_b=2, k=1). Unit feedback law means $n_f = n_g = 0$ (in equation 3.9), unit feedback also means $n_p = 0$ (in equation 3.9). Thus

$$\max [n_f - n_b, k+n_g - n_a -1] = -2 < n_p \qquad (3.19)$$

This means that the necessary and sufficient condition (equation 3.9) for closed-loop identifiability using parametric methods cannot be satisfied at all. The theoretical analysis indicates that no model of any structure representing the real process can be identified under pure feedback.

In Figure 3-3, the lowest plot corresponds to *S/N* = 0, only a straight line was identified, which reflected the inverse of the feedback controller. This is in accordance with the theoretical analysis in section 3.1.1.

Case 3

This case is to validate the theory on the influence of the noise model in closed-loop identification. In practice, no information on noise model is known *a priori*. A first-order noise model is presumed here, which is different from the true noise model and is not rich enough to catch the true noise characteristics. Through the direct approach (see subsection 2.1.3) and prediction error method with a Box-Jenkins model, the frequency responses and step response of the model identified under low signal to noise ratio (S/N = 0.410) and the real process are shown in Figure 3-4. In the same way, the frequency responses and step response of the model

identified under relative large signal to noise ratio (S/N = 3.685) and the real process are shown in Figure 3-5.

Noise model influence

As known from the process description (equation 3.18), the noise model of the true process is of second order. When a first-order noise model is assumed, the assumed noise model is quite different from the real process model. According to the theoretical analysis in section 3.1.1, there must be a bias in the identified process model. From Figure 3-4 and Figure 3-5, the bias (which is the difference between the true process model and the identified model) can be observed by comparing the solid line and the dashed line. If more carefully compared, the bias in Figure 3-5 is not so severe as that in Figure 3-4. That is because the estimate in Figure 3-4 is the result of small signal to noise ratio (*S/N* = 0.410) while the estimate in Figure 3-5 is the result of large signal to noise ratio (*S/N* = 3.685). The difference is due to different signal to noise ratio. This can also be explained from the equation (3.17). While signal to noise ratio is relatively large (*S/N* = 3.685), the first term is dominant. The identification tends to be open-loop identification. Thus the bias is relatively small.

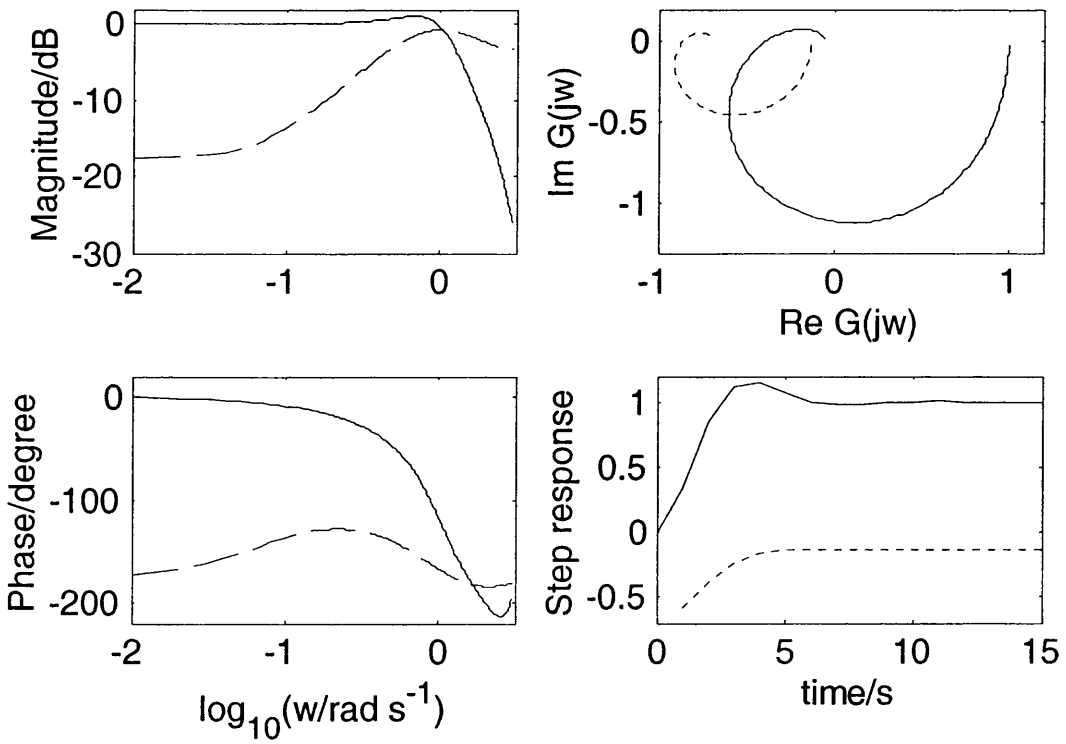


Figure 3-4 *Noise model influence under S/N=0.410: real process (solid line), the estimate (dashed line)*

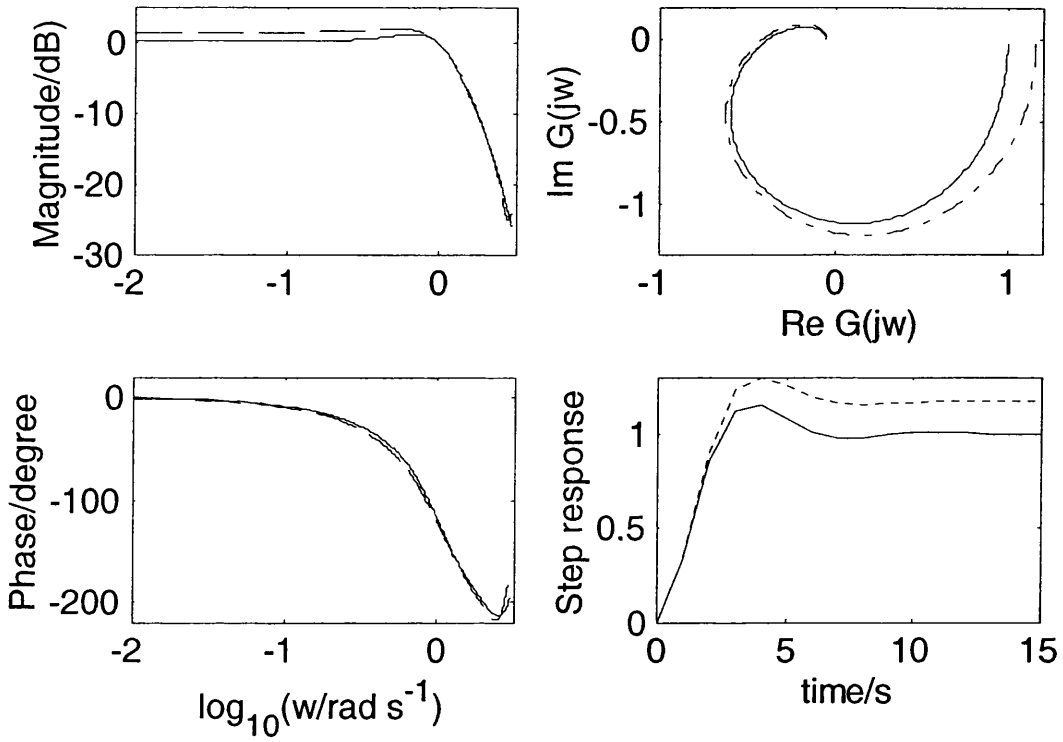


Figure 3-5 *Noise model influence under S/N = 3.685: real process (solid line), the estimate (dashed line)*

3.1.4 Conclusions

From the theory and simulation examples presented in this section, the conclusion is that:

- Under pure feedback, nonparametric methods can only identify the inverse of the feedback controller. When an external dither signal is injected, the estimate gives a weighting average of the real process frequency response and the frequency response of the inverse of the feedback controller. The weighting factor depends upon the signal to noise ratio.
- Under pure feedback, parametric models can only be identified if fairly restrictive necessary and sufficient conditions on the real process and controller orders are satisfied.
- Using direct approach, if the noise model is unknown and the assumed noise model is not rich enough to contain the true noise model, the estimate of the process will inevitably have a bias. When the signal to noise ratio is large, the bias will be small. Otherwise, when the signal to noise ratio is small, the bias will be large.

3.2 Closed-loop identification with external excitation

This section concentrates on analyses of the two-step method proposed by Huang and Shah (1997) and the two-stage method proposed by Van den Hof and Schrama (1993). Both of them are joint input-output approaches. The two methods will then be compared theoretically and by simulation. Conclusions will be given at the end.

3.2.1 Theory

In this subsection, theory on the two specific methods of joint input-output identification approach will be discussed in detail. Then the differences between Huang and Shah (1997) and Van den Hof and Schrama (1993) will be considered.

The two-step method (Huang and Shah 1997)

The closed-loop system in Figure 3-6, which is same as Figure 3-1 in subsection 3.1.1, is adapted.

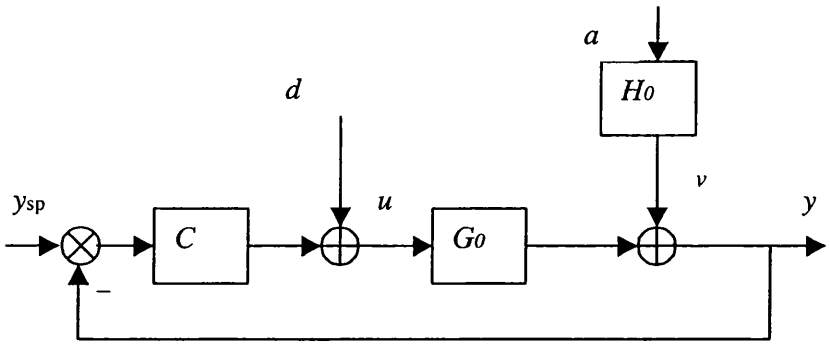


Figure 3-6 *Closed-loop system*

The equations (3.12) and (3.13) were rewritten as follows:

$$u(t) = S_0(q)d(t) - C(q)S_0(q)H_0(q)a(t) \qquad (3.20)$$

$$y(t) = G_0(q)S_0(q)d(t) + S_0(q)v(t) \qquad (3.21)$$

where the sensitivity function $S_0 = \dfrac{1}{1+G_0C}$ is defined in (3.11) that is the transfer function from disturbance v to process output y. These two equations (3.20 and 3.21) are the initiatives

of Huang and Shah (1997). Since *d(t)* and *a(t)* are uncorrelated signals, *u(t)* and *d(t)* are measurable, according to open loop identification theory (Ljung 1987), the sensitivity function $S_0(q)$ can be identified in an open loop way and can even be identified consistently. Once the estimate of sensitivity function is available, the output data are filtered with the inverse of the sensitivity function, equation (3.21) becomes

$$y(t)/S_0(q) = G_0(q)d(t) + H_0(q)a(t) \qquad (3.22)$$

Since *d(t)* and *a(t)* are uncorrelated, $G_0(q)$ can be consistently identified in an open loop way.

The identification steps in the Huang and Shah (1997) procedure are:

(1) The sensitivity function of the closed-loop system is identified.

(2) The inverse of the estimated sensitivity function is used to filter the output data for an open loop identification of the plant to be carried out.

The variance and bias considerations

Huang and Shah (1997) studied in detail the variance and bias of the closed loop identification and compared them with the corresponding open loop occasion. The conclusion is that " the presence of the sensitivity function is the key difference between open-loop and closed-loop identification" (Huang and Shah, 1997).

The two-stage method (Van den Hof and Schrama, 1993)

For Figure 3-6, the equation (3.12) has been derived in subsection 3.1.1 and is rewritten here for completeness

$$y(t) = G_0(q)S_0(q)d(t) + S_0(q)H_0(q)a(t) \qquad (3.23)$$

The equation (3.13) becomes

$$u(t) = S_0(q)d(t) - C(q)S_0(q)H_0(q)a(t) \qquad (3.24)$$

The equations (3.23) and (3.24) are the initiatives of Van den Hof and Schrama (1993). In (3.24), according to open loop identification theory (Ljung 1987), the sensitivity function

$S_0(q)$ can be identified in an open loop way and can even be identified consistently since *d(t)* and *a(t)* are uncorrelated.

From equation (3.24), it is assumed:

$$u^d(t) = S_0(q)d(t) \tag{3.25}$$

Signal $u^d(t)$ is a newly generated signal, which is a substitute of process input $u(t)$ to be used in the second step.

Thus equation (3.23) becomes

$$y(t) = G_0(q)u^d(t) + S_0(q)H_0(q)a(t) \tag{3.26}$$

In (3.26), $u^d(t)$ and *a(t)* are uncorrelated, so if $u^d(t)$ is measurable, $G_0(q)$ can also be identified in an open loop way and can be identified consistently according to open loop identification theory (Ljung 1987).

The identification steps in Van den Hof and Schrama (1993) are:

(1) The sensitivity function of the closed-loop system is identified using a high order FIR (finite impulse response) model.

(2) The estimated sensitivity function is used to simulate a noise-free input signal $u^d(t)$ for an open loop identification of the plant to be identified.

Differences between Huang and Shah (1997) and Van den Hof and Schrama (1993)

From the above description, both Huang and Shah (1997) and Van den Hof and Schrama (1993) are joint input-output approaches. They all have two steps in the identification procedures. The first step is the same, that is to identify the sensitivity function with process input *u(t)* and dither signal *d(t)*. The difference lies in the second step. Van den Hof and Schrama (1993) used the dither signal *d(t)* filtered by the sensitivity function to generate the simulated signal $u^d(t)$. Huang and Shah (1997) filtered the output *y(t)* with the inverse of the sensitivity function. These differences were out of different motivations. Huang and Shah (1997) intended to get the same accuracy in bias and variance and explicit expressions for

asymptotic variance and asymptotic bias. Van den Hof and Schrama (1993) aims at a consistent estimate of the process and an asymptotic bias expression only. (Huang and Shah, 1997)

3.2.2 Simulation example and case

To demonstrate the theoretical analysis and the methods discussed above, the same second order ARMAX model (as presented in subsection 3.1.2) is used (Huang and Shah, 1997). The transfer function is given by

$$y(t) = \frac{0.3403 + 0.2417q^{-1}}{1 - 0.7859q^{-1} + 0.3679q^{-2}} u(t-1) + \frac{1 - 0.8q^{-1} + 0.12q^{-2}}{1 - 0.7859q^{-1} + 0.3679q^{-2}} a(t) \tag{3.27}$$

As before, a unit feedback control law is implemented in this simulation, the disturbance $a(t)$ and the dither signal $d(t)$ are white noises and independent with each other. The number of data points is 5,000.

Case 4

The simulation is aimed at the comparison of the two methods of joint input-output approach (Huang and Shah, 1997; Van den Hof and Schrama, 1993) and the direct approach. In this case, the dither signal is kept σ_d^2 = 1 while the noise signal is kept σ_a^2 = 2.5 for all the three methods. No model structure and/or model order information is assumed to be known *a priori.*

3.2.3 Results and discussions

Case 4

The purpose of this case is to compare the two methods of joint input-output approach (Huang and Shah, 1997; Van den Hof and Schrama, 1993) and the direct approach. The condition for the following results is that assuming no model structure and/or model order information known *a priori.* Thus, whether model validation can pass or not becomes the sole proof for accepting the process estimate or not. For Huang and Shah (1997) and Van den Hof and Schrama (1993), the sensitivity function estimates were the intermediate results for the second step. Figure 3-7 shows the Bode plots of the real sensitivity function and the sensitivity function estimates. In Figure 3-8 and Figure 3-9, the real process and the estimates using the two methods are compared with the estimated result using the direct method in different way

(Bode plot, Nyquist plot and step response respectively). All the results were acquired with small signal to noise ratio as defined in Table 3.1 (i.e. $S/N = 0.410$).

From Figure 3-8 and Figure 3-9, the estimates using Huang and Shah (1997) and Van den Hof and Schrama (1993) are more accurate than that using the direct method. From Figure 3-8 and Figure 3-9 and Figure 3-7, the result from Huang and Shah (1997) is a bit better than that using Van den Hof and Schrama (1993).

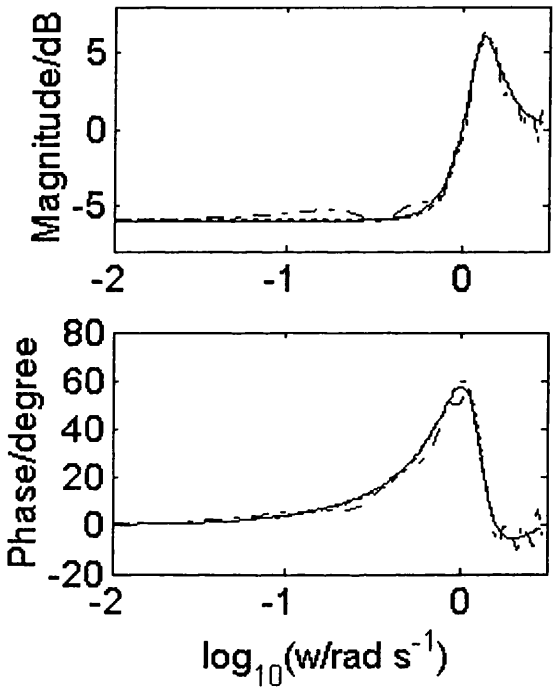


Figure 3-7 *Bode plots of the sensitivity functions: real sensitivity function (solid line), the estimate using Huang and Shah (1997) (dotted line), the estimate using Van den Hof and Schrama (1993) (dash dotted line)*

Huang and Shah (1997) and Van den Hof and Schrama (1993) belong to the same approach—joint input-output approach. Huang and Shah (1997) is a successful improvement of Van den Hof and Schrama (1993). The efforts of Huang and Shah (1997) and Van den Hof and Schrama (1993) represent a trend in closed-loop identification — using open-loop identification to solve the closed-loop identification problem. By comparison, Huang and Shah (1997) pay more attention to the comparison of the closed-loop identification with the open-loop identification. Van den Hof and Schrama (1993) may be more suitable for joint identification and control, an interesting topic in closed-loop identification.

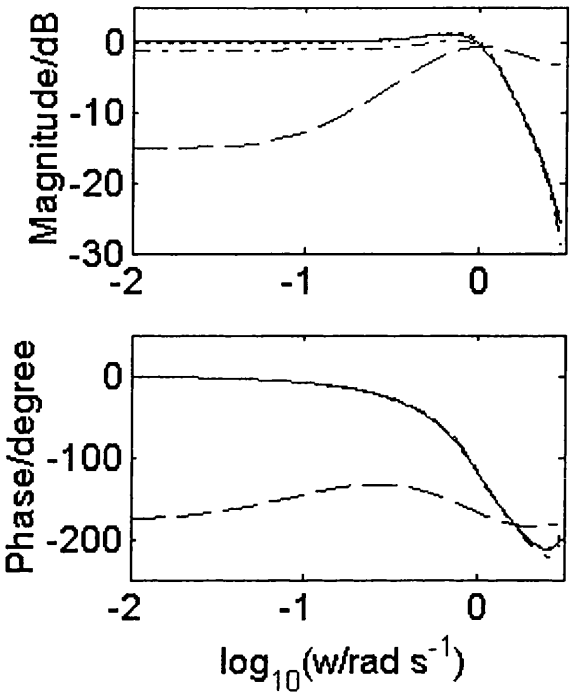


Figure 3-8 *Bode plots of the real process and the estimates: real process (solid line), the estimate using Huang and Shah (1997) (dotted line), the estimate using Van den Hof and Schrama (1993) (dash dotted line), the estimate with direct approach (dashed line)*

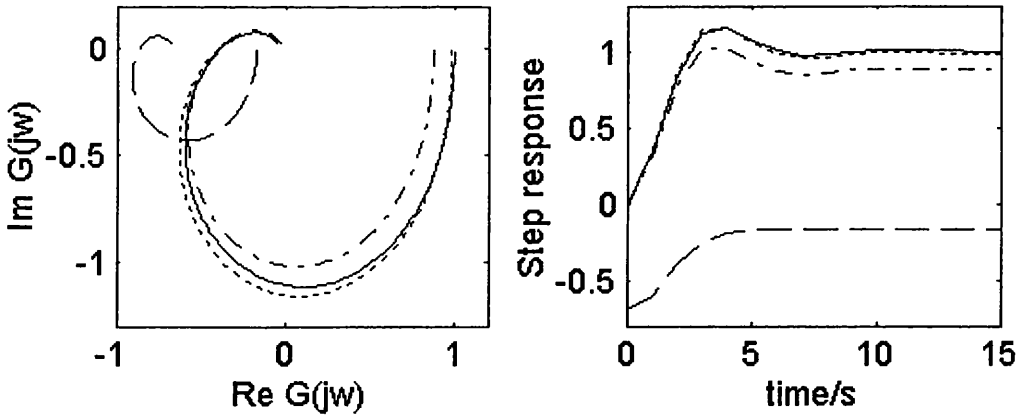


Figure 3-9 *Nyquist plots and step responses of the real process and the estimates: real process (solid line), the estimate using Huang and Shah (1997) (dotted line), the estimate using Van den Hof and Schrama (1993) (dash dotted line), the estimate with direct approach (dashed line)*

In both steps, Huang and Shah (1997) used BJ (Box-Jenkins) model structure, which includes information not only on process but also on noise. Van den Hof and Schrama (1993) used FIR model in the first step and OE model structure in the second step. Both FIR model and OE model assume a unit noise model. This can partly explain why the result from Huang and Shah (1997) is a bit better than that using Van den Hof and Schrama (1993). On the other hand, using BJ model (refer to subsection 1.2.2 for the introduction of BJ model) in Huang and Shah (1997) means more work in finding the suitable model order. This is a disadvantage.

3.2.5 Conclusions

Theoretical analysis and case study in this section leads to the following conclusion:

- Huang and Shah (1997) and Van den Hof and Schrama (1993) belong to joint input-output approach. There are two steps for both methods. The difference lies in the second step.
  - Van den Hof and Schrama (1993) used the dither signal *d(t)* filtered by the sensitivity function to generate the simulated signal $u^d(t)$.
  - Huang and Shah (1997) filtered the output *y(t)* with the inverse of the sensitivity function.

3.3 Relay-based identification

In this section, the basic theory of three relay-based identification methods is first introduced. Then two cases are designed to duplicate these three methods and to study the noise influence. Based on results from these cases, the advantages and disadvantages of these three methods are discussed. Conclusions are given in the last subsection.

For easy notation and clear comparison among the three relay-based identification methods, new names will be given to represent the above three methods in this section. It must be noted that the new name is not used in the literature. It only reflects the author's personal opinion.

- An Improved Autotune Identification Method: Li *et al.* (1991)
  In this thesis, it will be called "Two-test describing function method" with the acronym DF2 since there are two relay tests in the procedure and describing function analysis method is used in order to obtain a transfer function model.

- Low-order Modeling from Relay Feedback: Wang *et al.* (1997a)
  In this thesis, it will be called "One-test time domain method" with the acronym TD1 since there is one relay test and the derived waveform is analyzed with time-domain solution.

- Process Frequency Response Estimation from relay feedback: Wang *et al.* (1997b)
  It will be called "One-test FFT method" with the acronym FFT1 since it needs one relay test and FFT (Fast Fourier Transform) is used to analyze for results.

3.3.1 Theory

In this subsection, describing function analysis method will first be described in detail. Three different relay-based identification methods shown above will be analyzed. The issue of noise influence is then discussed because measurement noises are inevitable in industrial practice.

The idea of relay identification was to introduce a nonlinear feedback of the relay in order to generate a limit cycle oscillation (Figure 3-10). The system then starts to oscillate. The period and the amplitude of the oscillation are determined when steady-state oscillation is obtained (Figure 3-11). This gives the critical period and the critical gain (Åström and Hägglund, 1995).

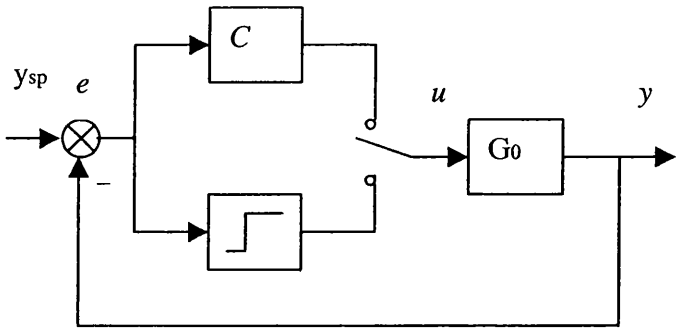


Figure 3-10 *Relay feedback system*

Describing function analysis

In order to analyze the system represented in Figure 3-10 which includes a relay, a nonlinear component, the concept of describing function analysis is introduced (see, for example, Atherton, 1975).

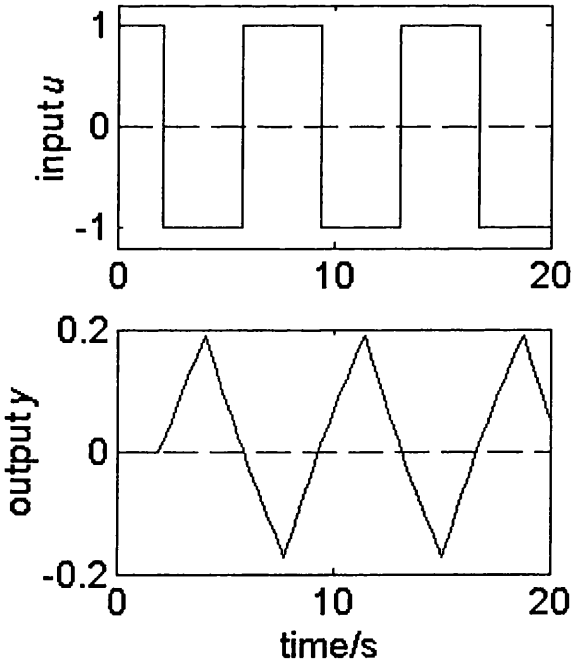


Figure 3-11 *Limit cycle from relay feedback*

A sinusoidal input $e(t)$, where a is the amplitude and ω is the frequency, is considered.

$$e(t) = a\sin(\omega t) \tag{3.28}$$

The relay output $u(t)$ in response to $e(t)$ would be a square wave as shown in Figure 3-11.

Using Fourier series expansion, it can be written as

$$u(t) = \frac{4d}{\pi}\sum_{i=1}^{\infty}\frac{\sin(2i-1)\omega t}{2i-1} \tag{3.29}$$

where d is the limit cycle amplitude, i is an integer.

If the process G behaves like a low pass filter and is responsive only to the fundamental Fourier harmonic of $u(t)$, i.e. the higher-frequency harmonics can all be neglected in the analysis, as compared with the fundamental component, the describing function $N(a)$ of the ideal relay is defined as the ratio of $u(t)$ and $e(t)$. That is

$$N(a) = \frac{4d}{\pi a} \tag{3.30}$$

The introduction of the describing function of the relay makes it possible for the system with non-linearity to be analysed as if it is a linear system. It should be noted that the describing function analysis method is an approximate method. The important assumption is that only the fundamental component of the relay output has been considered. This requires that the process G behave like a low pass filter.

If the process in Figure 3-10 is $G(j\omega)$, then for relay feedback system to oscillate continuously, the system must satisfy:

$$G(j\omega_u)N(a) + 1 = 0 \tag{3.31}$$

That is (Åström and Hägglund, 1995)

$$G(j\omega_u) = -\frac{1}{N(a)} = -\frac{\pi a}{4d} \tag{3.32}$$

In the Ziegler-Nichols frequency response method, the critical gain K_u is defined as

$$K_u G(j\omega_u) = -1 \tag{3.33}$$

So the critical gain for the system $G(j\omega)$ represented in Figure 3-10 is

$$K_u = \frac{4d}{\pi a} \tag{3.34}$$

At the same time, the critical frequency ω_u can be determined from the limit cycle by

$$\omega_u = \frac{2\pi}{P_u} \tag{3.35}$$

where P_u is the period of limit cycles.

DF2: An improved autotune identification method (Li *et al.*, 1991)

As reviewed in subsection 2.2.4, Li *et al.* (1991) improved from Luyben (1987). A modified procedure that does not require knowledge of the steady-state gain was proposed. The method includes two tests. The first is a normal one with a relay; the second is run with a relay plus an additional known dead time so that the phase angle is shifted about 45° and a smaller critical frequency is obtained. From equations (3.34) and (3.35) — the results from the describing function analysis method, there are two equations from each test. From the four equations, a least-square method is then used to determine the unknown parameters: time constant(s) and the steady-state gain.

The steps in the DF2 are:

(1) Critical gains (K_{u1} and K_{u2}) and critical frequencies (ω_1 and ω_2) are obtained from two tests. The additional dead time L_2 for the second test was chosen to make the additional

phase angle roughly $\pi/4$. This is because the frequency range from phase angle $-\frac{3}{4}\pi$ to $-\pi$ in Nyquist plot is of most interest.

(2) The unknown parameters for all three models are calculated from critical gains (K_{u1} and K_{u2}) and critical frequencies (ω_1 and ω_2) together with the dead times. The problem can be formulated as a linear least-squares problem.

(3) Comparing both the calculated parameters for each model and the residuals of the least-squares solutions, the best model is chosen. For an open-loop stable process, the correct model should have all positive parameters. If more than one model satisfies this condition, the model with smaller residual will be selected.

Some details will be described as following:

Step 1: From the first test, the period of the limit cycle P_{u1} and the amplitude of the limit cycle a_1 can be read out. According to (3.34) and (3.35), the critical frequency ω_1 and the critical gain K_{u1} are obtained. The additional dead time L_2 for the second test can be calculated using (3.36) (Li *et al.* 1991).

$$L_2 = \frac{5\pi}{12\omega_1} \tag{3.36}$$

From the second test, the period of the limit cycle P_{u2} and the amplitude of the limit cycle a_2 can be read out. The critical frequency ω_2 and the critical gain K_{u2} can be determined in the same way.

Step 2: The Least Square Calculation

Using model 1 to fit the data, that is

$$G(s) = \frac{Ke^{-Ls}}{Ts+1} \tag{3.37}$$

where K is the process steady-state gain, T is the time constant and L is the dead time (in this method, L is assumed to be read from the first test). For this model, the following equations are satisfied:

$$X_i = \frac{1}{K_{ui}}\cos(-\pi) = -\frac{1}{K_{ui}} \tag{3.38}$$

$$Y_i = \frac{1}{K_{ui}}\sin(-\pi) = 0 \tag{3.39}$$

where K_{ui} is the the critical gain of the i-th test (i = 1,2), that is K_{u1} and K_{u2} respectively.

$$z = [X_1 \quad Y_1 \quad X_2 \quad Y_2]^T \tag{3.40}$$

$$U = \begin{bmatrix} \omega_1 Y_1 & \cos(\omega_1 L) \\ -\omega_1 X_1 & -\sin(\omega_1 L) \\ \omega_2 Y_2 & \cos(\omega_2 L) \\ -\omega_2 X_2 & -\sin(\omega_2 L) \end{bmatrix} \tag{3.41}$$

Therefore, the unknown parameters can be obtained by

$$\hat{a} = \begin{bmatrix} T \\ K \end{bmatrix} = [U^T U]^{-1} U^T z \tag{3.42}$$

Here U will always have full rank since in (3.41) ω_1 represents the critical frequency and ω_2 represents the critical frequency when there is an additional delay. They are always slightly different.

Using model 2 (second-order model) and model 3 (third-order model) respectively, the parameters will be obtained in the same way (Li *et al.* 1991).

TD1: Low-order modeling from relay feedback (Wang *et al.*, 1997a)

Wang *et al.* (1997a) proposed an amplitude-biased relay feedback test. If an amplitude-biased relay (as shown in Figure 3-12) is introduced instead of an ideal relay in Figure 3-10, the resulting oscillation waveform of the process is shown in Figure 3-13. In Figure 3-12, u is the relay amplitude and μ_0 is a bias in the relay, ε is hysteresis width of the relay. In Figure 3-13, T_{u1} is the time period when the relay output is positive, T_{u2} is the time period when the relay output is negative, A_u is the positive peak of the limit cycle, A_d is the negative peak of the limit cycle.

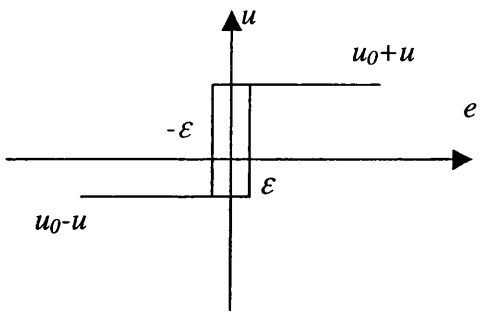


Figure 3-12 The *amplitude-biased relay*

A first-order plus dead-time (FOPDT) process below is assumed:

$$G(s) = \frac{Ke^{-Ls}}{Ts+1} \tag{3.43}$$

where K is process steady-state gain, T is the time constant and L is the dead time.

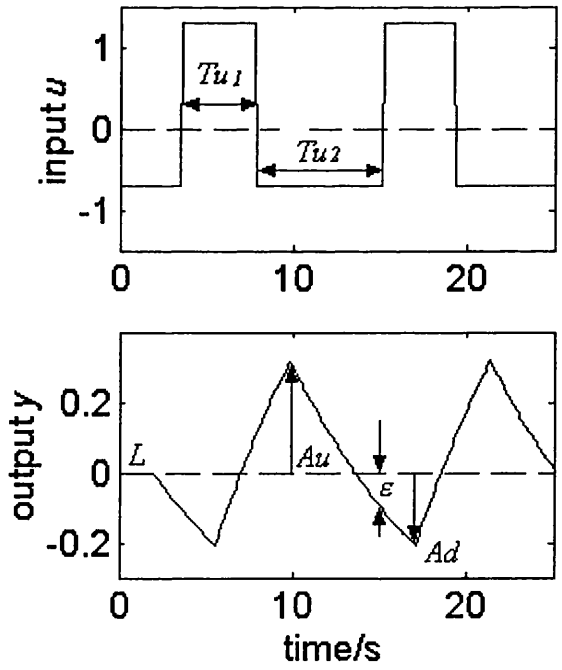


Figure 3-13 *Oscillatory waveforms under an amplitude-biased relay feedback for an FOPDT system*

After analyzing the limit cycle in time domain, Wang *et al.* (1997a) gave exact expressions of the period and the amplitude of limit cycle oscillation for the FOPDT (first-order plus dead-time) process (3.43). The process output *y(t)* converges to the stationary oscillation in one period ($T_{u1}+T_{u2}$) and the oscillation is characterized by

$$A_u = (\mu_0 + \mu)K(1-e^{-L/T}) + \varepsilon e^{-L/T} \tag{3.44}$$

$$A_d = (\mu_0 - \mu)K(1-e^{-L/T}) - \varepsilon e^{-L/T} \tag{3.45}$$

$$T_{u1} = T\ln\frac{2\mu K e^{L/T} + \mu_0 K - \mu K + \varepsilon}{\mu K + \mu_0 K - \varepsilon} \tag{3.46}$$

$$T_{u2} = T\ln\frac{2\mu K e^{L/T} - \mu K - \mu_0 K + \varepsilon}{\mu K + \mu_0 K - \varepsilon} \tag{3.47}$$

where μ is the relay amplitude, μ_0 is a bias in the relay and ε is hysteresis width of the relay (Figure 3-12).

Exact expressions for the periods and amplitudes of limit cycles have been obtained (as shown from 3.44 to 3.47). These four equations are sufficient to determine the three parameters of the process, but it is tedious to solve these equations. On the other hand, frequency response information contained in the asymmetric limit cycle of Figure 3-13 can be utilized. Since the waveforms of the process input $u(t)$ and the output $y(t)$ are periodic with the period ($T_{u1}+T_{u2}$), they can be expanded into Fourier series. The average components of these periodic waves can be extracted. Thus, the steady-state gain can be computed (Ramirez, 1985).

$$K = G(0) = \frac{\int_0^{T_{u1}+T_{u2}} y(t)dt}{\int_0^{T_{u1}+T_{u2}} u(t)dt} \tag{3.48}$$

With K known, the normalized dead time of the process $\theta = L/T$ can be obtained from (3.44) or (3.45) as

$$\theta = \ln \frac{(\mu_0 + \mu)K - \varepsilon}{(\mu_0 + \mu)K - A_\mu} \tag{3.49}$$

or

$$\theta = \ln \frac{(\mu - \mu_0)K - \varepsilon}{(\mu - \mu_0)K + A_d} \tag{3.50}$$

From (3.46) or (3.47),

$$T = T_{u1}(\ln \frac{2\mu K e^{\vartheta} + \mu_0 K - \mu K + \varepsilon}{\mu K + \mu_0 K - \varepsilon})^{-1} \tag{3.51}$$

or

$$T = T_{u2}(\ln \frac{2\mu K e^{\theta} - \mu_0 K - \mu K + \varepsilon}{\mu K - \mu_0 K - \varepsilon} \tag{3.52}$$

$$L = T\theta \tag{3.53}$$

Time-domain information is combined with frequency response point estimation so that a first-order plus dead time (FOPDT) model can be identified with a relay test.

The steps in the TD1 method —the one-test time domain method (Wang *et al.*, 1997a) are:

- The amplitude-biased relay experiment is performed. The process input $u(t)$ and output $y(t)$ are recorded, and the periods (T_{u1} and T_{u2}) and the amplitudes (A_u and A_d) of the oscillation are measured.
- Computation

  Step 1: Compute K from (3.48).

  Step 2: Compute θ from (3.49) or (3.50).

  Step 3: Compute T from (3.51) or (3.52).

  Step 4: Compute L from (3.53).

FFT1: Process frequency response estimation from relay feedback (Wang *et al.*, 1997b)

FFT-based Analysis

The describing function analysis method is only one of the limit cycle analysis methods. Another effective method is Fast Fourier Transform (FFT). If *y(t)* and *u(t)* in Figure 3-11 can be guaranteed to be absolutely integrable, then $Y(j\omega)$ and $U(j\omega)$ can be obtained from FFT of *y(t)* and *u(t)*. $G(j\omega)$ can be calculated from $Y(j\omega)/U(j\omega)$ (Ramirez, 1985).

From above, the prerequisite of using FFT to analyze the limit cycle is that *y(t)* and *u(t)* in Figure 3-11 be guaranteed to be absolutely integrable. An exponential decay $e^{-\alpha t}(\alpha>0)$ is introduced by Wang *et al.* (1997b) to overcome this problem. So

$$\tilde{y}(t) = y(t)e^{-\alpha t} \tag{3.54}$$

$$\tilde{u}(t) = u(t)e^{-\alpha t} \tag{3.55}$$

$\tilde{y}(t)$ and $\tilde{u}(t)$ are integrable because they will tend to zero after a period of time. Applying the Fourier Transform to (3.54) and (3.55) makes

$$\tilde{Y}(j\omega) = \int_0^\infty \tilde{y}(t)e^{-j\omega t}dt = \int_0^\infty y(t)e^{-\alpha t}e^{-j\omega t}dt = Y(j\omega+\alpha) \tag{3.56}$$

$$\tilde{U}(j\omega) = \int_0^\infty \tilde{u}(t)e^{-j\omega t}dt = \int_0^\infty u(t)e^{-\alpha t}e^{-j\omega t}dt = U(j\omega+\alpha) \tag{3.57}$$

Thus, the shifted process frequency response $G(j\omega_i+\alpha)$ can be calculated by

$$G(j\omega_i+\alpha) = \frac{Y(j\omega_i+\alpha)}{U(j\omega_i+\alpha)} = \frac{\tilde{Y}(j\omega_i)}{\tilde{U}(j\omega_i)} \tag{3.58}$$

To identify $G(j\omega)$, one can first take the inverse FFT of $G(j\omega_i+\alpha)$ as

$$\tilde{g}(kT) = FFT^{-1}(G(j\omega_i+\alpha)) = g(kT)e^{-\alpha kT} \tag{3.59}$$

Applying the FFT again to $g(kT)$ would result in the process frequency response $G(j\omega)$

$$G(j\omega_i) = FFT(g(kT)) \tag{3.60}$$

Noise issue

In industrial practice, there will be measurement noise. To demonstrate the relay-based identification method practically, it is necessary to discuss how to prevent measurement noise influence and to explore the actual noise influence.

A hysteresis in the relay is one way to reduce the measurement noise influence. The hysteresis width should be larger than the noise band (Åström and Hägglund, 1995). Generally, hysteresis width is chosen two times larger than the noise band (Wang *et al.*, 1997b).

Filtering is another way to prevent noise influence (Åström and Hagglünd, 1984a). A lowpass analogue Butterworth filter is usually used because the measurement noise is of high frequency. The cut-off frequency of the Butterworth filter can be determined with respect to the process frequency region of interest and is usually chosen to be 3 to 5 times of the process critical frequency (Wang *et al.*, 1997b).

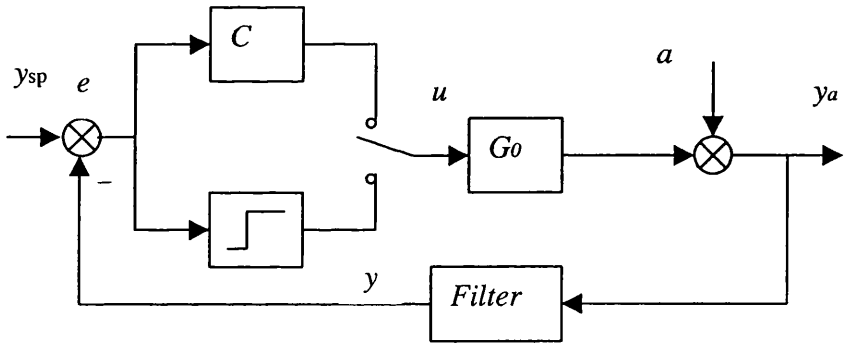


Figure 3-14 *Relay feedback system with Butterworth filter (a is a time-continuous white noise)*

There are different ways to use the Butterworth filter for eliminating the noise influence. One way is to put the filter after the sensor (in series in the main loop) as shown in Figure 3-14. After the process (plus the filter) has been identified, the filter can be removed from the estimated results. Another way is to use two Butterworth filters, one filter in the relay output and one in the process output. Both filters are not in the main feedback loop. The filter will be cancelled when the process is identified.

With the first way (i.e. the filter after the sensor — in series in the main loop), the filter can remove most of the noise. Therefore, the relay hesteresis width needed is very small. On the contrary, with the second way, it is not necessary to remove the filter from the estimated result. But, the relay hesteresis width needed will be much larger. In this thesis, the first way is employed for further study.

For a good understanding of online analogue Butterworth filter design and implementation in relay-based identification, more technical details on analogue Butterworth filter design will be presented here. The frequency response of a typical Butterworth filter is shown in Figure 3-15 (Krauss *et al.*, 1994).

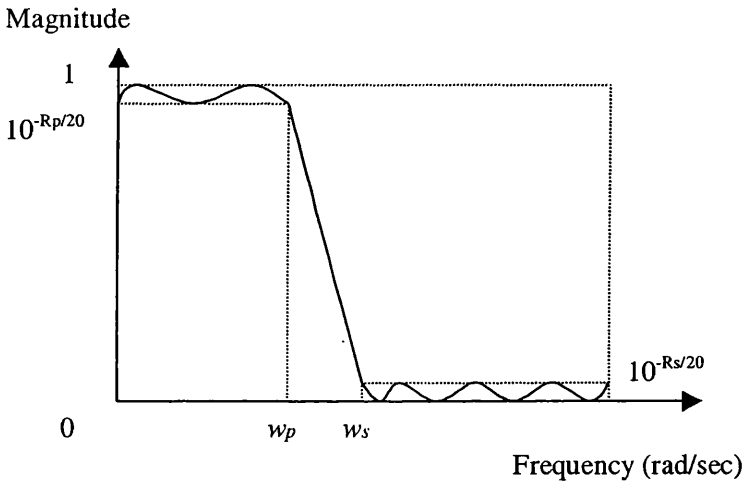


Figure 3-15 *Frequency response of a typical Butterworth filter (the solid line), the dashed line shows the limits.*

Some parameters have to be determined beforehand in order to design an analogue Butterworth Filter. These parameters are:

ω_p : passband corner frequency or the cut-off frequency (in rad/sec)

ω_s : stopband corner frequency (in rad/sec)

R_p : passband ripple (in decibels or dB)

R_s : stopband attenuation (in decibels or dB)

$\omega_s - \omega_p$: frequency transition width (in rad/sec)

The physical meanings of these parameters are shown in Figure 3-15.

3.3.2 Simulation Examples and Cases

The transfer function simulated was

$$G_0 = \frac{e^{-2s}}{10s+1} \tag{3.61}$$

This is the most commonly used example in relay-based identification literature (e.g., Li *et al.*, 1991, Wang *et al.*, 1997b and Shen *et al.*, 1996b) and therefore allows comparisons.

Case 5

The purpose of Case 5 is to duplicate the three methods discussed above. Simulations without noise were done. For the DF2 method (Li *et al.*, 1991) and the FFT1 method (Wang *et al.*, 1997b), the height of the relay is set up to be 1 (Figure 3-10). No hysteresis was added to the relay because no noise was applied. For the TD1 method (Wang *et al.*, 1997a), the biased relay shown in Figure 3-12 is: relay height μ is 1 and the relay bias μ_0 is 0.3, hysteresis width ε is 0.1.

Case 6

The purpose of Case 6 is to see the noise influence in relay-based identification when both measures (relay hysteresis and filtering) in the end of Subsection 3.3.1 were taken. The FFT1 method (Wang *et al.*, 1997b) was selected.

An online analogue Butterworth lowpass filter is used as well as the hysteresis (with hysteresis width 0.1) (as shown in Figure 3-14). For the process (3.61) with time constant 10 seconds and time delay 2 seconds, its critical frequency ω_c is about 0.8 red/sec. Therefore, the cut-off frequency of the Butterworth filter ω_p is chosen to be 2.4 rad/sec (i.e. 3 times the process critical frequency ω_c) and the stopband corner frequency ω_s is chosen to be 4.0 rad/sec (i.e. 5 times the process critical frequency ω_c) because the Butterworth filter can not be very steep; passband ripple R_p is 3 dB and stopband attenuation R_s is 15 dB (Figure 3-15). The Butterworth filter obtained is a 4-order s function

$$\frac{46.2613}{s^4 + 6.815s^3 + 23.222s^2 + 46.3525s + 46.2613}$$

For the FFT method, relay height μ used is 1 and hysteresis width ε is 0.1 (Figure 3-12). The applied white noise (with variance 0.1) is injected with the Band-Limited white noise module in Matlab/Simulink (The MathWorks; Natwick, MA). Two tests were designed with the same noise. The only difference is relay height (either 1 or 1.5). For objective comparison, the same online analogue Butterworth lowpass filter is used as well as the hysteresis (with hysteresis width 0.1) for both tests.

3.3.3 Results

The results of the above simulation cases are presented in this subsection.

Case 5

This case is aimed at duplicating the three methods (DF2, TD1 and FFT1) introduced above under the condition of no noise.

DF2: An Improved Autotune Identification Method (Li et al., 1991)

After simulating with the first-order plus dead time (FOPDT) process (3.61), the process input and output for two tests were plotted as shown in Figure 3-16.

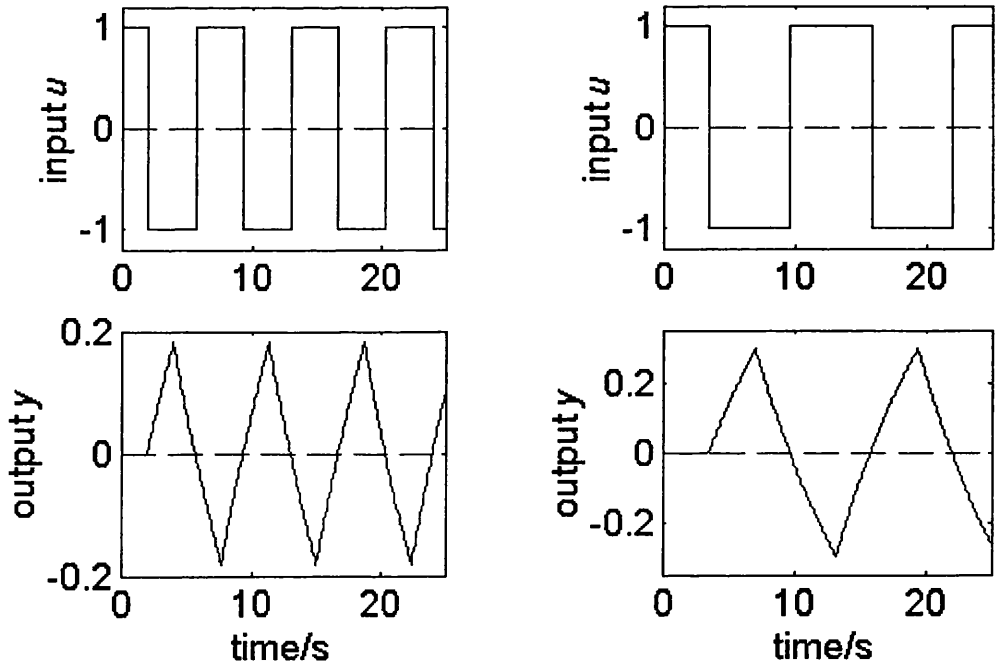


Figure 3-16 *Time trends for the DF2 method, the first test (left); the second test (right)*

From the first test, *the dead time L was read first*; the period of limit cycle P_{u1} and the amplitude a_1 of principal harmonic of the output were also read out, thus the critical frequency ω_1, critical gain K_{u1} and additional dead time for the second test L_2 can be calculated

$$L = 2.000 \qquad P_{u1} = 7.335 \qquad a_1 = 0.181$$

From the second autotune test, the following results can be obtained.

$$P_{u2} = 12.282 \quad a_2 = 0.289$$

Using the DF2 method, the estimated model is

$$\hat{G} = \frac{0.976e^{-2s}}{8.042s + 1} \tag{3.62}$$

The result published in Li *et al.* (1991) is

$$\hat{G} = \frac{0.988e^{-2s}}{8.020s + 1} \tag{3.63}$$

Table 3-2 described the results duplicated *vs* the results published in Li. *et al.* (1991).

Parameters	L_2	K_{u1}	ω_1	K_{u2}	ω_2
From Case Study	1.530	7.023	0.857	4.407	0.512
Published	1.526	7.031	0.858	4.287	0.513

Table 3-2 *Results duplicated vs results published in Li. et al. (1991)*

The Nyquist plots of the real process and the model obtained by using the DF2 method are given in Figure 3-17. The conclusion from the above results is that the method has been implemented correctly.

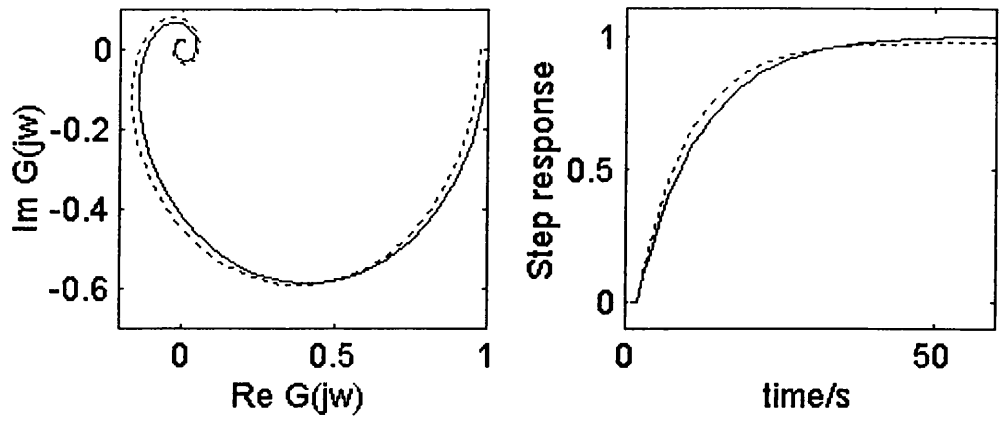


Figure 3-17 *Nyquist plots of the real process and the estimate obtained by using the DF2 method (Li et al., 1991)*

TD1: Low-order Modeling from Relay Feedback (Wang et al., 1997a)

Using the same process (3.61) and the biased relay (Figure 3-12), the time trends were logged from the simulation as shown in Figure 3-18.

From Figure 3-18, the following measurements can be obtained.

$$A_u = 0.317 \qquad A_d = -0.208$$

$$T_{u1} = 4.291 \qquad T_{u2} = 7.278$$

Using the TD1 method, the estimated model is

$$\hat{G} = \frac{1.007e^{-1.998s}}{10.084s + 1} \tag{3.64}$$

The Nyquist plots of the real process and the model obtained by using the TD1 method are given in Figure 3-19. The uncertatinty analysis of the estimated model is presented in ***Appendix B***.

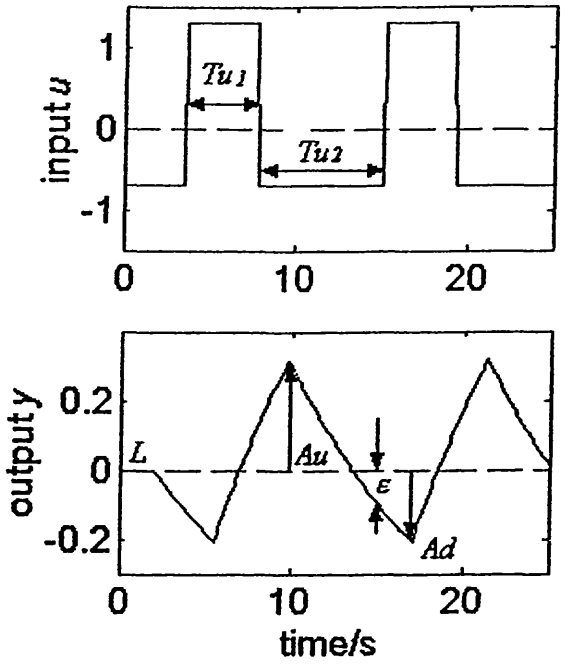


Figure 3-18 *Time trends by using the TD1 method*

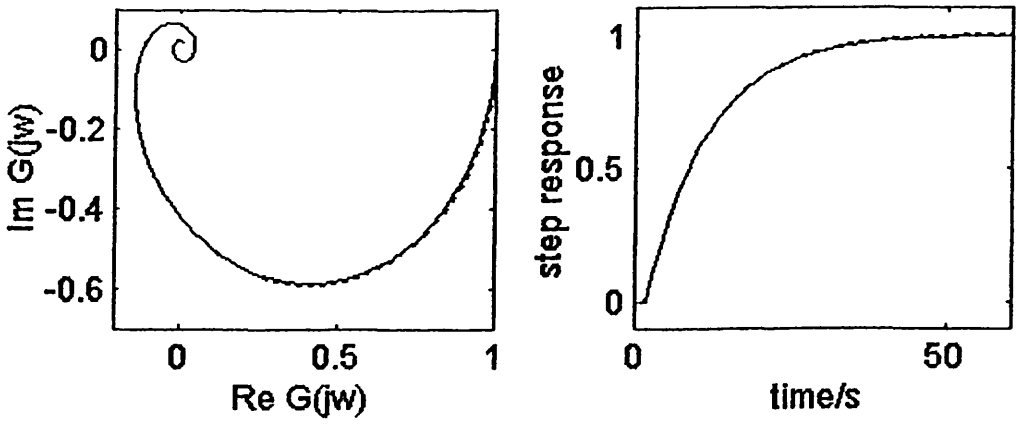


Figure 3-19 *Nyquist plots of the real process and the estimate obtained by using the TD1 method (Wang et al., 1997a): real process (solid line), estimate (dot line)*

Remark: Based on extensive simulations, a remarkable criterion to accept or reject the identification result by using the TD1 method is to calculate θ with formula (3.49) and (3.50)

respectively. If the difference of the two values of θ is greater than 0.01, then reject the result; if both θ have the same value, it is the best result!

FFT1: Process frequency response estimation from relay feedback (Wang et al., 1997b)

With the same process (3.61), the process output and input were logged from simulation as shown in Figure 3-20. It should be noted that in the actual duplication, the sampling rate was 0.05 second, the decay exponential used was 0.01535 (i.e. $e^{-\alpha}$ with $\alpha = 0.01535$) and the total simulation time was 600 seconds. If this were used for Figure 3-20, then it would be too dense to see anything meaningful. Here for effective presentation, the sampling rate was 0.1 second, the decay exponential used was 0.09 and the total simulation time was 100 seconds.

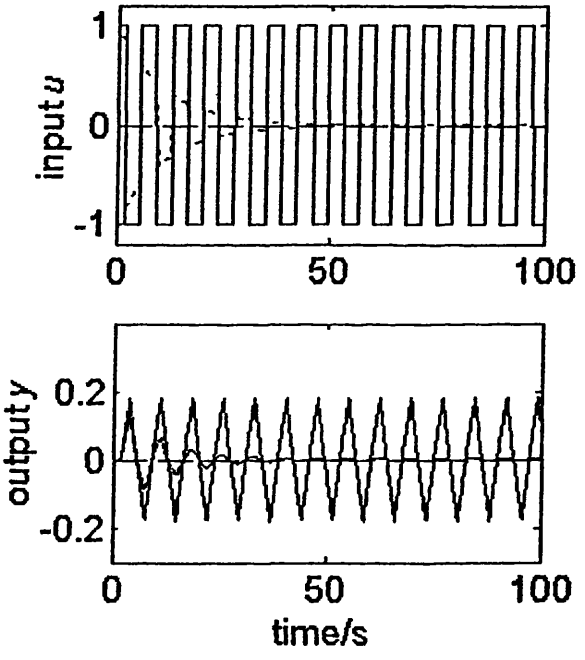


Figure 3-20 *Time trends by using the FFT1 method, real input u and output y (Solid line); the corresponding waveform after the exponebtial decay was introduced (dotted line)*

Using the FFT1 method, the result obtained is a Nyquist plot instead of a transfer function model. The Nyquist plots of the real process and the model obtained from the FFT1 method are given in Figure 3-21.

Remark: Result in Figure 3-21 indicates the FFT1 method has been duplicated successfully. The only difference is sampling rate. In Wang et al. (1997b), the sampling rate reported is 0.12 second; while the sampling rate used for this case study is 0.05 second.

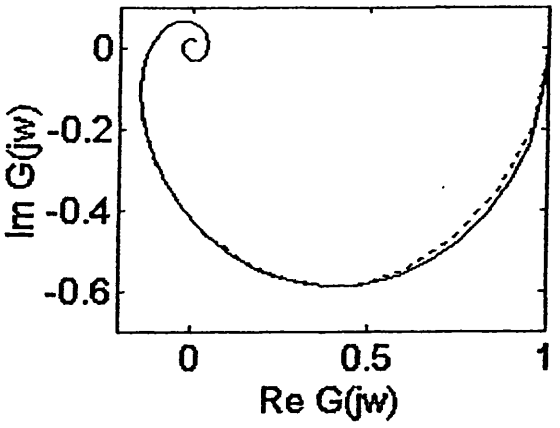


Figure 3-21 *Nyquist plots of the real process and the model obtained from the FFT1 method (Wang et al., 1997b): real process (solid line), estimate (dot line)*

Comparison of Accuracy with the three methods discussed

The Nyquist plots of the estimates obtained by using the DF2 method, the TD1 method and the FFT1 method respectively are compared in Figure 3-22. The accuracy of the three methods can be seen qualitatively.

To see the accuracy quantitatively, the definition of Worst Case Error is first introduced (this will also be introduce in Subsection 5.3.1 and widely used in Chapter 6). For continuous-time, it is

$$ERR\% = \left| \frac{\hat{G}(j\omega) - G(j\omega)}{G(j\omega)} \right| \times 100\% \qquad (3.65)$$

For discrete-time frequency response, the definition is

$$ERR\% = \left| \frac{\hat{G}(j\omega_i) - G(j\omega_i)}{G(j\omega_i)} \right| \times 100\% \quad i = 1,2,\cdots M \qquad (3.66)$$

where $G(j\omega_i)$ and $\hat{G}(j\omega_i)$ are the actual and the estimated process frequency response respectively. Only the Nyquist curve from zero frequency to the crossover frequency ω_{cg} where the argument is -180° were considered since this part is the most important for system identification and controller design.

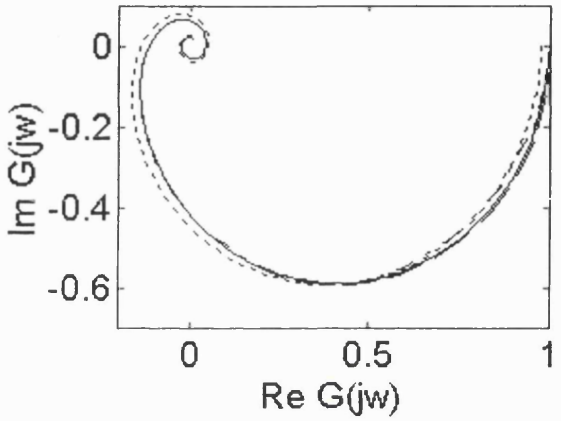


Figure 3-22 *Nyquist plots of the three estimates with the real process: real process (solid line), the DF2 method (dot line), the TD1method (dashed line), the FFT1 method (dash-dot line)*

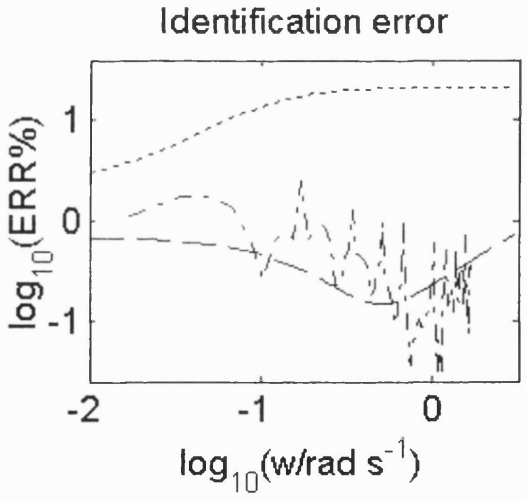


Figure 3-23 *Identification errors of the three methods:the DF2 method (dotted line), the TD1 method (dashed line) ,the FFT1 method (dash-dot line)*

For this FOPDT process (3.61), the critical frequency ω is between 0.8 *rad/sec* and 1 *rad/sec*. Because of this, in identification error plot (Figure 3-23), only frequencies ranging from ω=0.1 rad/sec to ω=5 rad/sec were discussed.

From Figure 3-23, in case of the accuracy, the TD1 method is the best, the DF2 method is the worst while the FFT1 method is slightly worse than the TD1 method, but much better the DF2 method.

Case 6

This case is to investigate noise influence in relay-based identification. The FFT1 method was chosen. The two different tests have the same noise (variance 0.1), but with different relay height (either 1 or 0.5). The time trend from simulation with relay height 1 is shown in Figure 3-24.

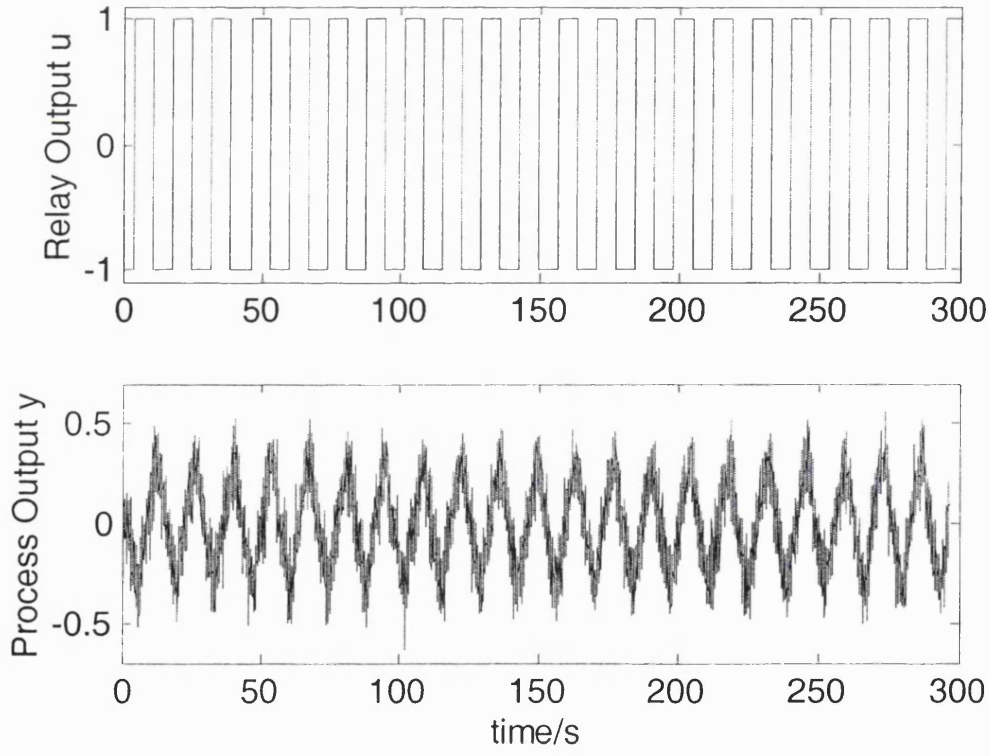


Figure 3-24 *Time trends for the test with white noise(variance 0.1) and relay height 1.*

From Figure 3-25, both of the tests have good identified results. By the comparison of the lower two panels, the accuracy of the estimate with relay height 0.5 is not as good as that with relay height 1.0.

In relay identification, low-pass filtering and relay with hysteresis are used to prevent noise influence. The low-pass filter can deal with noise above the cutoff frequency (in the frequency domain) effectively. For the noise beyond this region, it still relies on the exciting signal power to prevent noise influence. The relay height is proportional to the signal energy injected into the controlled loop. The higher the relay height, the more the signal power. With the same noise, the case relay height 1.0 means higher signal-to-noise ratio than the case relay height 0.5.

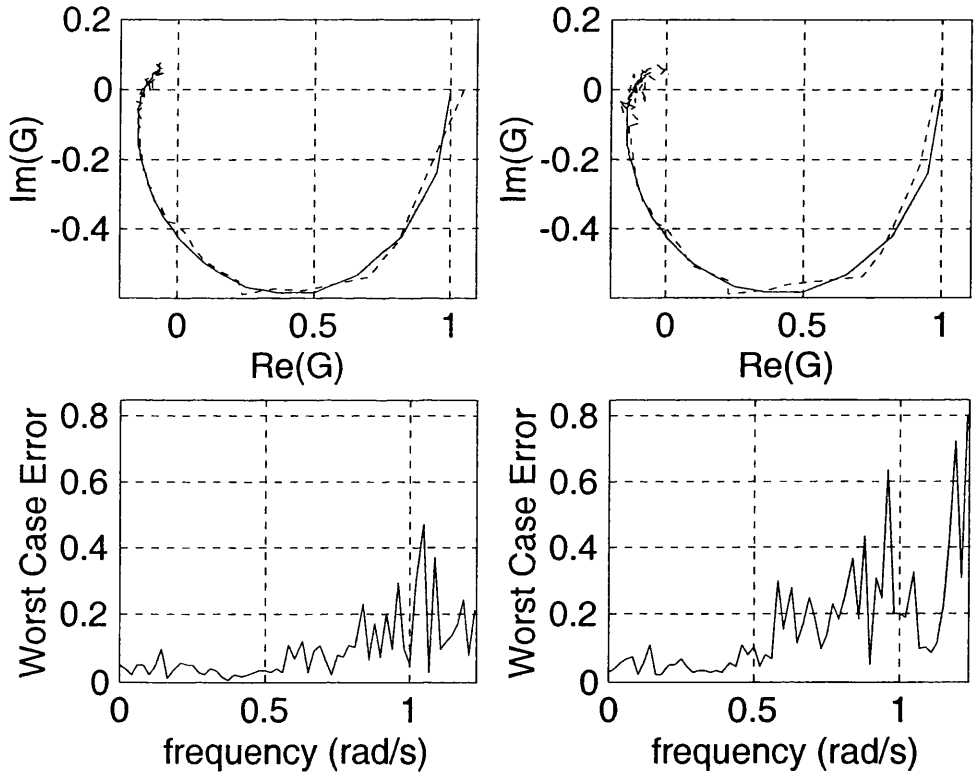


Figure 3-25 *Measurement Noise Influence of the FFT1 method (Wang et al. 1997b): real process (solid line), estimate measurement noise (dotted line). The left two panels for relay height 1; the right two panels for relay height 0.5.*

3.3.4 Discussions

In this subsection, the results identified by using the three relay-based identification methods and the cost to obtain these models will first be discussed. Then the efforts will be directed to find out the reasons. The methods and its acronyms will be rewritten here for clarity.

DF2: An Improved Autotune Identification Method (Li et al., 1991)
TD1: Low-order Modeling from Relay Feedback (Wang et al., 1997a)
FFT1: Process frequency response estimation from relay feedback (Wang et al., 1997b)

Identification Result

- *Information contained in the model*

  From the frequency point of view, the DF2 method identifies the information at frequency points $\omega = \omega_u$ and approximately $\angle G(j\omega) = -\frac{3\pi}{4}$. The TD1 method identifies the information at frequency points $\omega = \omega_u$ and $\omega = 0$. Thus the above two methods are known as two-point methods. The FFT1 method identifies the information at multiple frequency points (the number determined by researcher according to one's purpose, generally, the information lie in $0 \sim \omega_u$). This can be observed from Figure 3-26.

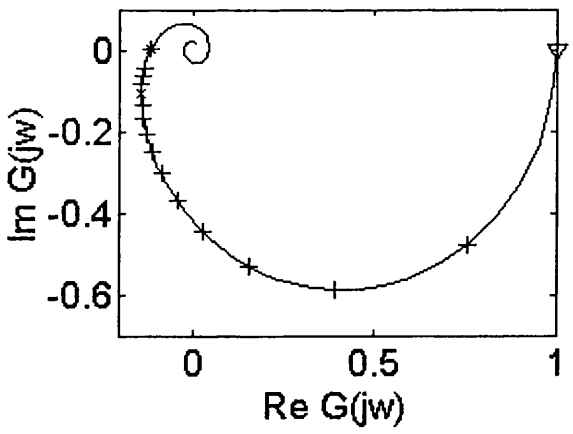


Figure 3-26 *Information identified by different methods. The DF2 method identifies the critical frequency (*) and one point in the third quadrant (×). The TD1 method identifies the critical frequency (*) and the zero frequency (∇). TheFFT1 identifies all the points in this figure.*

- *Accuracy*

  For the first-order plus dead time (FOPDT) process (3.61), the TD1 method is the most accurate, the DF2 method the least. The FFT1 method is slightly worse than the TD1 method, but much better than the DF2 method (Figure 3-23). The reasons behind the above phenomena lie in: (1) the DF2 method is based on describing function analysis. When the critical frequency and critical gain are computed from the limit cycles, only the primary harmonic component is considered. In essence, this is an approximate analysis method. (2) the TD1 method is based on exact analysis of the period and amplitude of limit cycle in the time domain. No approximation is made in this derivation. Unfortunately, the time-domain analysis was done by only assuming the process is FOPDT. For more complex process than that in equation (3.61), the TD1 method has the possibility to degrade in accuracy because the method is based on the assumption that the true process is FOPDT (first order plus dead time). (3) the FFT1 method uses Fast Fourier Transform. No approximation is made and no model information is needed *a priori* to obtain the process estimate.

- *Influence of the Measurement noise*

  With regard to the three relay-based identification methods discussed, if no further steps are taken, they are prone to be influenced by measurement noise. Relay with hysteresis and filtering are two ways to prevent noise influence. The low-pass filter can deal with noise above the cutoff frequency (in the frequency domain) effectively. For the noise from zero to the cutoff frequency (in the frequency domain), the exciting signal power is important in preventing noise influence. The relay height is proportional to the signal energy injected into the controlled loop. The higher the relay height, the more the signal power.

Cost

Here cost means what has to be done to let the method work and obtain the result.

- *Test number and relay type*

  The FFT1 method needs only one test. The TD1 method also needs one test. The DF2 method needs two tests and the set-up of the second test depends on the result of the first one.

  An ideal relay is used in the FFT1 method. The TD1 method requires the relay be an amplitude-biased relay. The amplitude-biased relay influences the process more severely

than a relay since it will cause the operating point to drift. A relay is used in each test in the DF2 method and an additional time delay is needed.

- *Computation*

  In case of computation needed to get the model, the FFT1 method the costliest, the DF2 method the next, the TD1 method the least. The following facts help explain why the above happen. (1) In the FFT1 method, the FFT transform has to done several times. For good accuracy and appropriate frequency scale, the sampling interval must be small enough, the sample numbers must be large enough. All these lead to much more computation. (2) the DF2 method has to try different model structures and then to compare the coefficients (all positive) and the residuals (the least) to find the model. (3) the TD1 method combines the time-domain information with the frequency response point estimation. This greatly simplifies the computation.

3.3.5 Conclusions

From the theoretical analysis and the cases studied in this section, the conclusions are:

- The FFT1 method and the TD1 method are much more accurate than the DF2 method since the describing function analysis is used in the DF2 method. The TD1 method has the potential to degrade in accuracy for processes that are more complex than FOPDT (first order plus dead time).
- The FFT1 method and the TD1 method need only one test. The DF2 method needs two tests. The relay in the TD1 method has to be amplitude-biased which will cause the operating point to drift.
- Relay-based identification methods are prone to be influenced by measurement noise. Two measures are effective in preventing noise influence, one is a relay with hysteresis; the other is to insert on-line Butterworth filter in the main loop.

3.4 Comparison of the two categories of CLID

In this section, the issues associated with the comparison of the two identification methods are analyzed first. Then two simulation cases were designed for the comparison. The results of the comparison between the closed-loop identification with external excitation (e.g. Huang and Shah, 1997) and the relay-based identification (e.g. the FFT method by Wang *et al.* 1997b) are subsequently presented. The conclusion is given at the end.

3.4.1 Issues for comparison and solutions

Issues

To compare the relay-based identification and identification with external excitation by simulation, the following two issues have to be faced.

- *Conversion between discrete-time description to continuous-time description*

  In relay-based identification, the real process is generally given in continuous-time form, with measurement noise (often relatively small). However, in closed-loop identification with external excitation, the real process is generally given in discrete-time form, with white noise (often relatively large). On the other hand, relay-based identification generally collects point information (such as peak values and zero-crossing points) from simulation to derive a continuous-time transfer function, while identification with external excitation is often done by applying time-series analysis to discrete-time transfer function models. The first problem with the comparison is that there exists no undisputed algorithm for parameter translation from discrete-time parameters to a continuous-time description and at which sampling rate/time for parameter translation from discrete-time parameters to a continuous-time description.

- *Definition of signal to noise ratio*

  In relay-based identification, the excitation is a relay. In closed-loop identification with external excitation, the excitation is a persistently exciting signal (for example, a white noise). The different excitations have different probability distributions (the former one is Binomial distribution; the latter is Normal distribution as shown in Figure 3-27). How to measure their power is an open question. More important are the different definitions of signal to noise ratio in two identification categories. In relay-based identification, the signal to noise ratio is defined as the power of the process output divided by the power of the

noise. In closed-loop identification with external excitation, the signal to noise ratio is defined as the power of the freshly injected signal divided by the power of the noise. The difference lies in the interpretation of the signal.

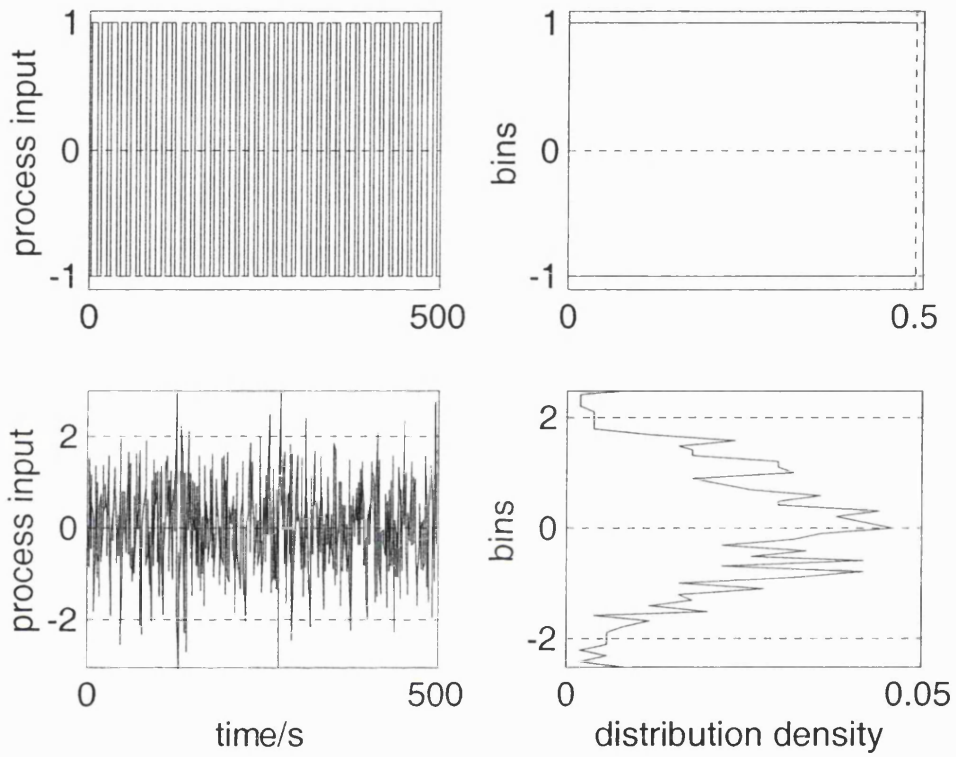


Figure 3-27 *Time trend and probability density function for the process inputs: the top two panels for the relay-based identification; the bottom two panels for the closed-loop identification with external excitation.*

Solutions

- To avoid the conversion between discrete-time description to continuous-time description, a continuous transfer function model can be used as the simulation process. For different methods to work, different sampling rates could be used.
- For comparison of results, the Nyquist plots of the identified models can be used. This will avoid the conversion of identified models from continuous to discrete or from discrete to continuous time.
- For the excitation signal power, the Binomial distribution for the relay can be calculated (Figure 3-10 in p84):

$$\text{var} = E(r - \bar{r})^2 \qquad (3.67)$$

$$\sigma_r = \sqrt{E(r-\bar{r})^2} \qquad (3.68)$$

where r is the relay signal. Generally, $\bar{r} = 0$ since the relay output is symmetrical. If d is used to express the relay amplitude, then $\sigma_r = d$.

The Normal distribution of white noise can be calculated (Figure 3-6 in p76)

$$\sigma_d = \sqrt{\text{var}(d(t))} \qquad (3.69)$$

Therefore, the relay amplitude may be chosen so that $\sigma_d = \sigma_r$.

- The definition of signal to noise ratio which will be used in the forthcoming simulation examples is:

$$S/N = \frac{\sigma_u^2}{\sigma_a^2} \qquad (3.70)$$

where σ_u^2 and σ_a^2 are the variance of the process input u and the noise a respectively.

3.4.2 Simulation Examples and Cases

From the analysis of the issues for comparison, strictly speaking, they cannot be compared. But for some qualitative comparison conclusions, the following simulation examples and cases were designed.

The FOPDT model as shown in 3.61 is again chosen as the simulation process since this process is commonly used as simulation example. Different sampling rates (0.05 second and 1 second respectively) were used for the relay-based identification (e.g. the FFT method by Wang *et al.* 1997b) and the closed-loop identification with external excitation (e.g. Huang and Shah, 1997) to work. Band-limited White Noise module in Matlab/Simulink is used as noise source (with default seed, sample time 0.05, the noise power 0.005) for both methods. The variance of the noise is 0.1003.

$$G_0 = \frac{e^{-2s}}{10s+1} \qquad (3.61)$$

Case 7

For relay-based identification with the FFT method (Wang *et al.* 1997b), the relay amplitude a used is 1; the energy injected into the process is $\sigma_u^2=1$, noise signal $\sigma_d^2=0.1003$. Therefore, the signal to noise ratio is 9.970.

For closed-loop identification with external excitation (Huang and Shah, 1997) to identify the process (3.61) under the same condition as in Wang *et al.* (1997b), a random number module in Matlab simulink with variance 0.855 is employed which caused the process input variance $\sigma_u^2=1$. Therfore, this simulation has nearly the same signal to noise ratio.

Case 8

The same band-limited white noise module in Matlab/Simulink is used as noise source with noise variance 0.1003. In this case, the signal to noise ratio is greatly reduced.

For relay-based identification with the FFT method (Wang *et al.* 1997b), the relay amplitude a used is 0.50; the energy injected into the process is $\sigma_u^2=0.25$, noise signal $\sigma_d^2=0.1003$. Therefore, the signal noise ratio is 2.493.

For closed-loop identification with external excitation (Huang and Shah, 1997) to identify the process (3.61) under the same condition as in Wang *et al.* (1997b), a random number module in Matlab/Simulink with variance 0.143 is employed which caused the process input variance $\sigma_u^2=0.25$. Therefore, this simulation, closed-loop identification with external excitation (Huang and Shah, 1997) has the same signal to noise ratio as relay-based identification (Wang *et al.*, 1997b).

3.4.3 Results

The real processes and the estimates are compared in Figure 3-28. The worst-case errors for the estimates resulting from the different methods are also presented in the same figure.

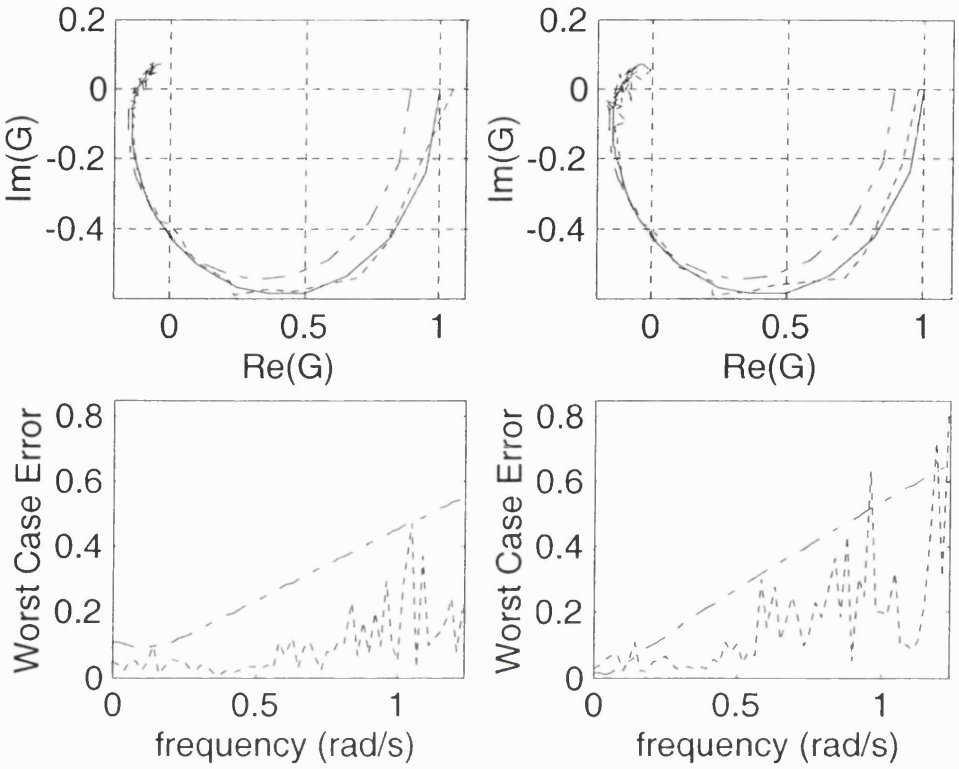


Figure 3-28 *Comparison of identification results and the worst case error; the left two panels for Case 7 where* $\sigma_u^2 = 1$*; the right two panels for Case 8 where* $\sigma_u^2 = 0.25$*; in the four panels, solid line for true process, dotted line for relay-based identification result, dashdot line for closed-loop identification with external excitation result.*

3.4.4 Discussions

From the above simulation cases and identification results, the qualitative conclusions of the comparison are as follows:

- *A priori information of the identified system*

  In the relay-based identification, the critical frequency of the identified system can be automatically found without *a priori* information of the identified system. In closed-loop identification with external excitation, *a priori* information on the critical frequency of the identified system is needed for the design of the persistently exciting signal (signal bandwidth). If this information is not available, an improper exciting signal could lead to a very bad identified result.

- *Accuracy and Noise Influence*

  From Figure 3-28, both categories of methods can achieve accurate result. Generally, the relay-based identification (e.g. Wang *et al.* 1997a) needs higher signal to noise ratio. When signal to noise ratio is reduced, the accuracy resulting from the relay-based identification will decrease sharply.

  The reason behind the phenomena is the different mechanisms used to deal with noise. The relay-based identification deals with noise by low-pass filtering and hysteresis as detailed at the end of subsection 3.3.1. The low-pass filter can prevent noise in the frequency range above cutoff frequency effectively (for example, the cut-off frequency of the Butterworth filter ω_p is chosen to be 3 times ω_c in subsection 3.3.2). In the frequency range from 0 to the cut-off frequency, the accuracy depends on the signal to noise ratio. In closed-loop identification with external excitation, the noise is dealt with by choosing an appropriate noise model structure and noise model order. Whenever the noise model is correctly estimated, the identified result can be accurate even for very low signal to noise ratios.

3.4.5 Conclusions

In this section, the two categories of CLID method have been compared with a typical method in each category.

The conclusions are:

- In the relay-based identification, the critical frequency of the identified system can be automatically found. In closed-loop identification with external excitation, *a priori* information on the critical frequency of the identified system is needed for the design of the persistently exciting signal.
- The relay-based identification method generally requires higher signal to noise ratio for good accuracy. The closed-loop identification with external excitation method can give accurate estimates even for small signal to noise ratio. This is due to the different mechanisms in each method used to deal with noise.

3.5 Conclusions

This section summarises the comparison of the relay-based identification and identification with external excitation. The following conclusions can be drawn from the above four sections. The comparison mainly involves the system components, the identifiability condition, identification results and the identification costs. The main advantages and disadvantages are given in the end.

Feedback system components:

- *Excitation*

  In closed-loop identification with external excitation, a fresh injection (generally persistently exciting) is often needed (*d* in Figure 3-6). In relay-based identification, generally the excitation into the loop is an ideal relay or an amplitude-biased relay with or without a hysteresis (Figure 3-10 and Figure 3-12).

- *Disturbance/noise*

  In closed-loop identification with external excitation, the noise is often assumed to be white noise. In relay-based identification, the measurement noise is often assumed to be white noise and the load disturbance is often assumed to be step like even though load disturbance is not discussed in this chapter.

- *Process domain*

  In identification with external excitation, the process is always given in discrete-time form, for example, the ARMAX process in equation 3.27. In relay-based identification, the process is mostly represented in continuous-time form, for example, an FOPDT (first-order plus dead time) process in equation 3.61.

Identifiability

- *Identifiability*

  In closed-loop identification with external excitation, a sufficient and necessary condition (3.9) should be met for parametric methods under pure feedback. Otherwise, if there is an external excitation signal, identifiability will not be a problem. In relay-based identification,

the problem has seldom been discussed except by Åström and Hagglünd (1984b). The key issue being the generation of stable limit cycles.

Results

- *Information obtained (frequency-domain)*

  In closed-loop identification with external excitation, information of the process over a range of frequencies can be identified. In relay-based identification, generally only one point (the critical point) can be identified even though there are methods for two or multiple point identification.

- *Accuracy and noise influence*

  The relay-based identification method generally requires higher signal to noise ratio for good accuracy. The closed-loop identification with external excitation method can obtain accurate estimates even in small signal to noise ratio.

Cost

Cost is used to mean what is required for the method to work and for acceptable accuracy.

- *Information needed*

  In closed-loop identification with external excitation, it is necessary to know the critical frequency approximately. Otherwise, the bandwidth of the exciting signal may not be suitable for obtaining good estimate. In relay-based identification, almost no *a priori* information is needed. The relay height and/or hysteresis width can be determined by experiment.

Main advantages and disadvantages

- *Closed-loop identification with external excitation*

  The main advantage for closed-loop identification with external excitation is that information of the process over a range of frequencies can be identified and that accurate estimates can be obtained even for small signal to noise ratio. The main disadvantage for closed-loop identification with external excitation is that *a priori* information on the critical frequency of the identified system is needed for the design of the persistently exciting signal.

- *Relay-based identification*

The main advantage for relay-based identification is that the sustained oscillation can be implemented automatically. Therefore, no *a priori* information on the critical frequency of the identified system is needed. The main disadvantage for relay-based identification is that the controller must be replaced by an ideal relay, which means disruption to the control system.

CHAPTER 4

THEORY ON CLID WITH A QUANTIZER

This chapter aims to develop new closed-loop identification theory. It begins by presenting the study of various quantizers because the quantizers play a key role in signal generating for the new closed-loop identification scheme. Then in the second section, the characteristics of quantization error (with the change of quantization interval) are analysed in order to gain the insight and to provide basis for further discussion. Methods that can be used corresponding to different ranges of quantization interval are subsequently presented. Key issues that will be encountered in the CLID methods are analysed. Conclusions are given at the end.

4.1 Study of the Quantizer

In this section, the difference between sampling and quantization are first clarified. The special quantizers used by Goodwin and Welsh (1999) and Welsh and Goodwin (1999), as well as the quantizer in Matlab Simulink (The MathWorks; Natwick, MA), are evaluated. The quantizer used in this thesis is then described.

4.1.1 Quantization *vs* Sampling

Before discussion of the quantizer in more detail, it is necessary to point out the difference between quantizing and sampling.

Quantizing is the process of mapping the samples into a sequence of binary digits that represent the quantized amplitude of the signal (Proakis and Manolakis, 1988). The function of quantizing is implemented by a quantizer. A quantizer is a non-linear component that provides a many-to-one mapping.

In the context of control, *sampling* means replacing a continuous-time signal by a sequence of numbers, which represents the values of the signal at certain times. The times when the measured signals are converted to digital form are called the sampling instants. The time between successive samplings is called the sampling period (Åström and Wittenmark, 1997).

Remark: One important difference is that sampling is the operation in the time axis while quantization is the operation in the amplitude axis.

In most cases, sampling occurs together with quantization. For example, when a digital control system is used, a process variable is sampled and then quantized since the process is continuous-time, i.e. the A/D conversion process involves sampling and quantizing.

Figure 4-1 illustrates the above differences. A sine wave with period 4 seconds (top left panel of Figure 4-1) is a continuous analogue signal. If the original signal is sampled with sample rate 0.2 second, a sampled analogue signal is obtained (top right panel of Figure 4-1). If the original signal is quantized with quantization interval 0.5, a continuous quantized signal is obtained (bottom left panel of Figure 4-1). If the original signal is sampled with sample rate 0.2 second and is then quantized with quantization interval 0.5, a sampled quantized signal is obtained (bottom right panel of Figure 4-1).

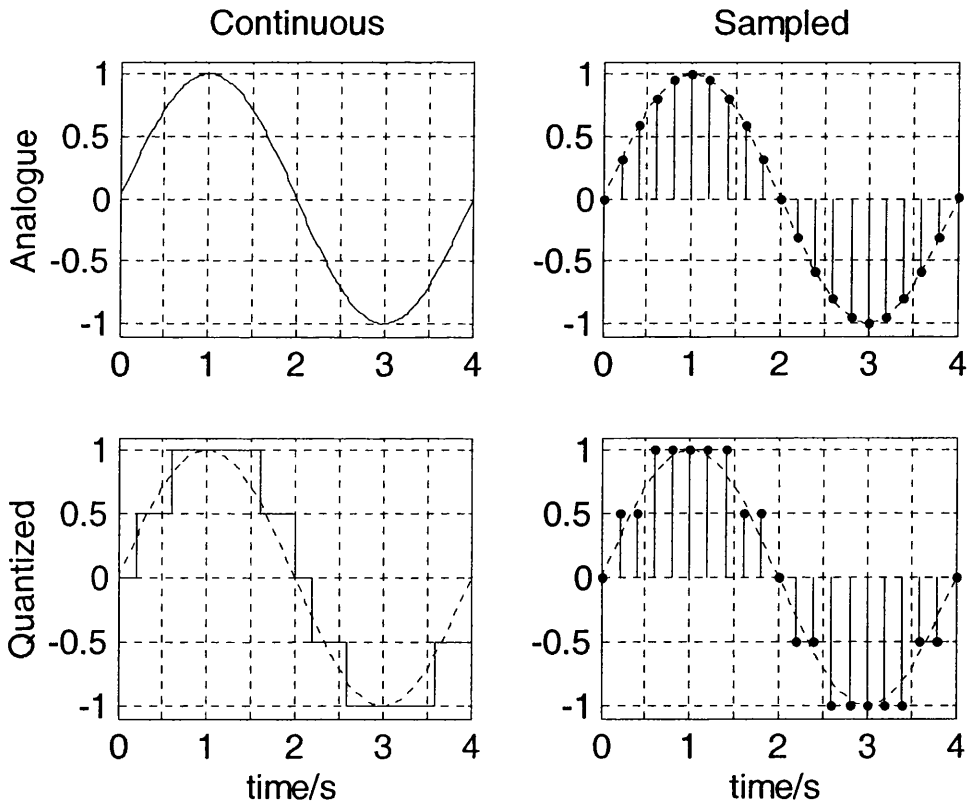


Figure 4-1 *Continuous/sampled, analogue/quantized signals; the left two panels are continuous signals, the right two panels are sampled signals; the top two panels are analogue signals, the bottom two panels are quantized signals.*

4.1.2 Special Quantizers

Goodwin and Welsh (1999) and Welsh and Goodwin (1999) defined two special quantizers called mid-rise quantizer and mid-step quantizer.

> The mid-rise quantizer allows the loop critical frequency to be determined essentially as in the relay experiments. The mid-step quantizer modifies the controller to have nonlinear characteristics so that small loop gains apply to small amplitude signals whilst normal loop gains are retained for dealing with large disturbances.
>
> (Goodwin and Welsh, 1999:3347)

The mid-rise quantizer

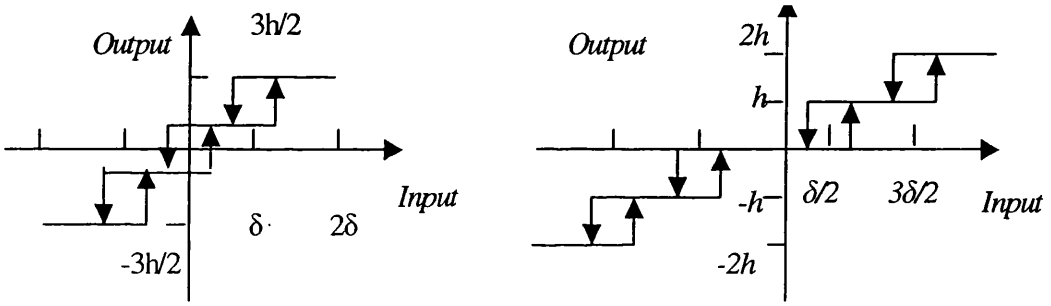


Figure 4-2 *The mid-rise quantizer(left) and the mid-step quantizer(right)*

The mid-rise quantizer is shown in Figure 4-2 (left panel). Parameters for the mid-rise quantizer include quantization interval (*qi*), quantizer amplitude (h) and hysteresis width (ε).

Welsh and Goodwin (1999) studied the analogy between the relay and the mid-rise quantizer and pointed out that the quantizer can be viewed as a set of relays operating at different input levels with respect to their output level.

Remark: The mid-rise quantizer behaves like a relay with hysteresis. The difference is that the mid-rise quantizer has multiple output levels while the relay has only two output levels.

The mid-step quantizer

The mid-step quantizer is illustrated in Figure 4-2 (right panel). Parameters for the mid-step quantizer also include quantization interval (*qi*), quantizer amplitude (h) and hysteresis width (ε).

Remark: The difference between the mid-step quantizer and the mid-rise quantizer is that zero is a quantization level for the former and zero is always in the middle of two quantization levels for the latter.

The quantizer in Matlab Simulink

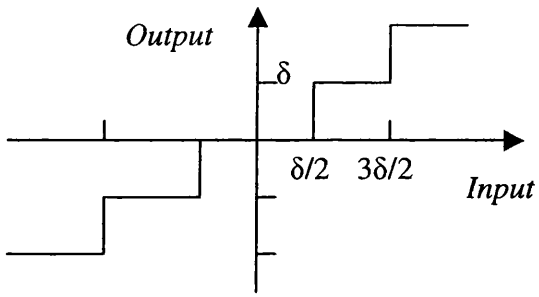


Figure 4-3 *The quantizer in Matlab Simulink*

The output-input relationship for the quantizer in Matlab Simulink (The MathWorks; Natwick, MA) is shown in Figure 4-3. The quantizer in Matlab Simulink has only one parameter - the quantization interval (*qi*). The quantizer amplitude (h) is implicitly assumed to be equal to the quantization interval (*qi*).

Remark: If there is no hysteresis and the quantization interval (*qi*) is equal to h, the mid-step quantizer behaves the same as the quantizer in Matlab Simulink.

4.1.3 Quantizer used in this thesis

The three kinds of quantizers discussed above can be generated as special cases of a quantizer with four parameters: quantization level (*ql*), quantization interval (*qi*), quantizer amplitude (h) and hysteresis width (ε). Because it is common practice to assume that the quantization interval (*qi*) is equal to the quantizer amplitude (h), the newly-defined quantizer will be termed as a three-parameter quantizer, i.e. quantization level (*ql*), quantization interval (*qi*) and hysteresis width (ε). The quantizer implemented with three parameters (as shown in Figure 4-4) is used in this thesis for further work.

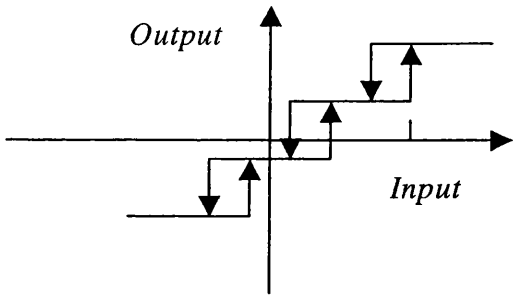


Figure 4-4 *The quantizer used in this thesis*

For the three-parameter quantizer, the special cases are discussed below. If zero is a quantization level *(ql)* and the hysteresis width (ε) is not zero, then the quantizer is a mid-step quantizer. If zero is a quantization level *(ql)* and the hysteresis width (ε) is zero, then, the quantizer is the quantizer in Matlab Simulink. If zero is in the middle of two quantization levels *(ql)* and the hysteresis width (ε) is not zero, then, the quantizer is a mid-rise quantizer.

The three-parameter implementation is practical and flexible for industry since the setpoint of a control loop is somewhere between 4mA and 20mA instead of 0mA in real practice. This will be shown in Chapter 5 and Chapter 6 where the three-parameter quantizer will be used repeatedly for simulation and experimentation.

4.2 Analysis of the quantization error

In this section, an example with a unit amplitude sine wave passed through the quantizer in Matlab Simulink (The MathWorks; Natwick, MA) is presented first. Then the characteristics of the quantization error for the two occasions are examined. The conclusion on the relationship of information content changing with the quantization interval is given at the end.

4.2.1 An example to study the quantization error

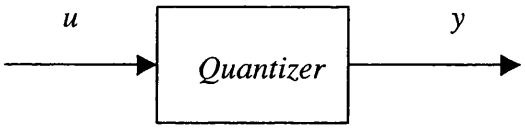


Figure 4-5 *the example for quantization error study*

As shown in Figure 4-5, u is a unit amplitude sine wave; the quantizer used is the quantizer in Matlab Simulink. The quantizer input u, quantizer output y and the quantization error (y-u) were plotted in Figure 4-6. In the left panel, the quantization interval (qi) is 0.2 and in the right panel, the quantization interval (qi) is 0.5.

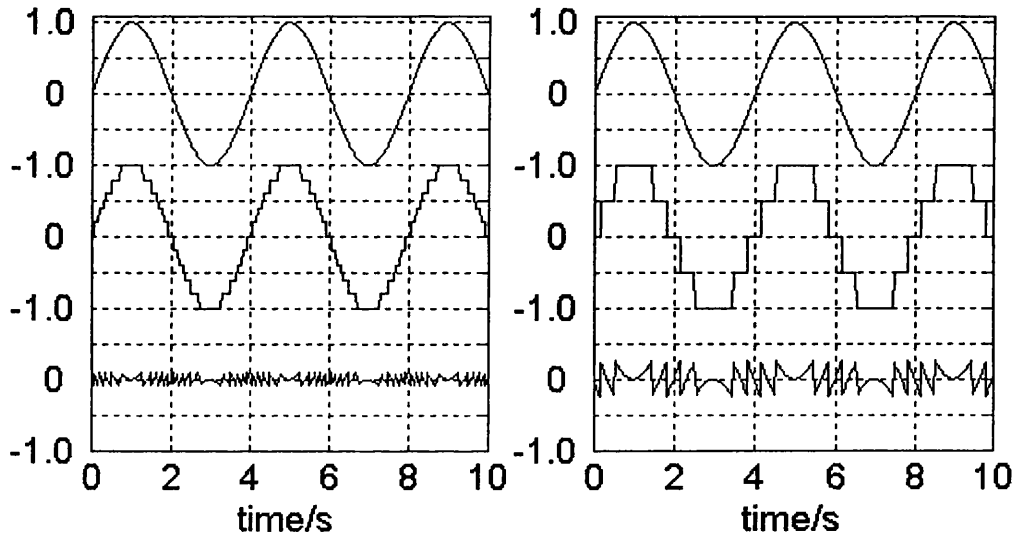


Figure 4-6 *The unquantized signal (top), quantized signal (middle) and the quantization error (bottom)*

4.2.2 Spectral characteristics of the quantization error

The information content of a signal includes power and bandwidth. In this subsection, the relationship between the quantization interval and the information content of the quantization error will be studied through the two cases described in subsection 4.2.1 (i.e. a unit sine wave is quantized with $qi = 0.2$ and $qi = 0.5$ respectively).

The spectrum is the square of the magnitude of the Fourier transform. The spectrum can be calculated with a finite transform known as discrete Fourier transform (DFT). The DFT is

usually computed using the fast Fourier transform (FFT). For the quantization error signals plotted in Figure 4-6, the corresponding spectral characteristics were plotted in Figure 4-7. In the algorithm to plot the power spectrum, a Hanning window was used.

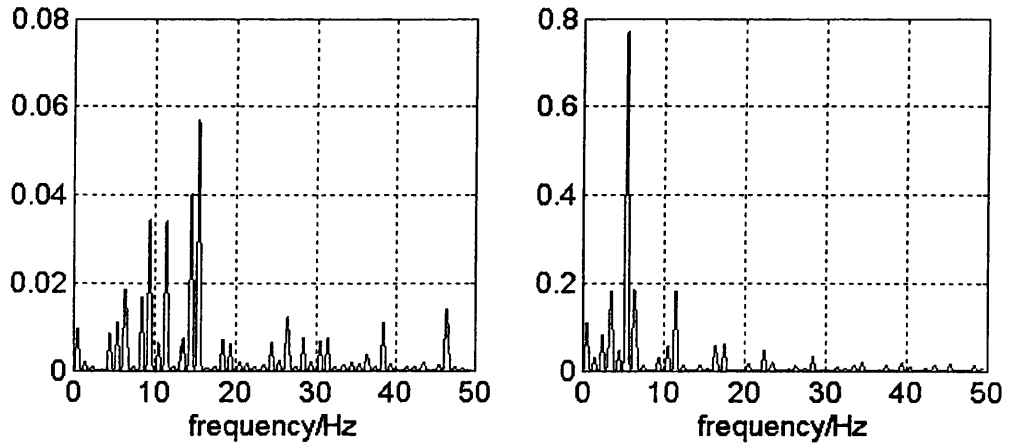


Figure 4-7 *Spectral characteristics corresponding to the quantization error plotted in Figure 4-6; In the left panel, the quantization interval (qi) is 0.2 and in the right panel, the quantization interval (qi) is 0.5.*

From Figure 4-7, in the left panel, the quantization interval is 0.2, the bandwidth of the quantization error is wide, that is to say, it has spectral features across the frequency range. The power of the quantization error is small. However, in the right panel, the quantization interval is 0.5, the bandwidth of the quantization error is not so wide because most of the power is concentrated below about 12 Hz, but the power of the quantization error is much higher by comparison.

Remark: When the quantization interval (*qi*) changes from small to large, the power of the quantization error will vary from small to large, while the bandwidth of the quantization error will vary from wide to narrow.

Therefore, from the study of the above two cases, a trade-off exists between these two elements. The above conclusion is important because it provides a theoretical basis for Section 4.3 below.

4.2.3 Nonlinearity of the quantization error

This section investigates the change of nonlinearity index for different quantiztion intervals. A linear time series has a dynamic model such as the Box Jenkins model with constant

coefficients driven by Gaussian white noise (Ljung 1987). In contrast, a nonlinear signal cannot be expressed by any model with constant coefficients. The purpose of the nonlinearity index test is to find why the quantization error signal can be used for closed loop identification.

The principle for nonlinearity test is described in subsection 5.3.2. Table 4.1 is the nonlinearity test result for the data sets discussed above (i.e. quantization interval 0.2 and 0.5 respectively).

	unquantized	Quantized
Quantization interval 0.2	1.15	1.25
Quantization interval 0.5	1.15	1.36

Table 4.1 *Nonlinearity index for unquantized and quantized signals*

Remark: When the quantization interval (*qi*) changes from small to large, the nonlinearity of the quantization error will vary from small to large.

4.3 Theory on CLID based on quantization

In relay-based identification, the controller has to be switched out of the control loop to let the relay in. This means a disruption to the control loop. In Welsh and Goodwin (1999) and Goodwin and Welsh (1999), a special quantizer (mid-rise or mid-step) was placed at the output of the controller removing the need to switch off the controller.

Welsh and Goodwin (1999) studied the analogy between the relay and the mid-rise quantizer in the context of relay-based identification. The conclusion is that the mid-rise quantizer can be viewed as a set of relays operating at different input levels with respect to their output level. The mid-rise quantizer has advantages and disadvantages relative to the relay. The principal advantage is that the mid-rise quantizer maintains control of the system if a large disturbance occurs at the plant. The disadvantage of the mid-rise quantizer is related to the open loop gain of the system, which may switch to a new level and a more complex form of analysis will be required. The relay has only two states and cannot switch to another level and therefore will not amplify the signal further.

In this section, new theory will be presented and different methods will be employed for CLID based on quantization. A quantizer will be inserted in the feedback path as shown in Figure 4-8. It will be shown in this section that different CLID methods can be used

corresponding to different quantization interval. The main advantage of the proposed scheme is that it does not require the injection of external excitation. The quantizer provides excitations on both occasions. In contrast, Welsh and Goodwin (1999) requires excitation at a specific frequency to be injected into the control loop to aid the identification.

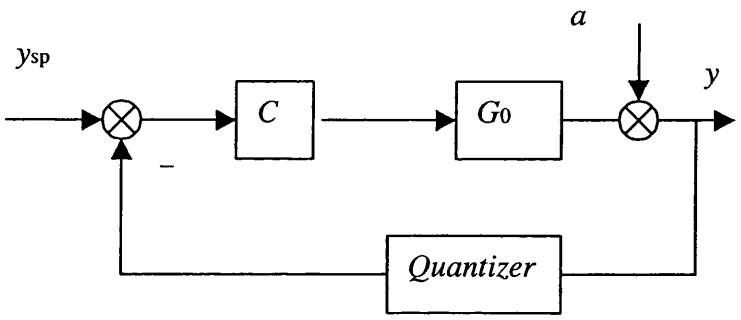


Figure 4-8 *Quantizer inserted in the feedback path.* G_0 *is the true process,* y *is the measurement,* a *is the measurement noise.*

4.3.1 The quantization interval continuum from small to large

From the study of the characteristics of the quantization error in section 4.2, the following observation can be made. When the quantization interval is small, there must be a range of different quantization interval values that make the quantizer error excitation eqivalent to a persistently exciting external signal; when the quantization interval is large, nonlinearity will be dominant, the quantizer is equivalent to a relay. This is shown in Figure 4-9.

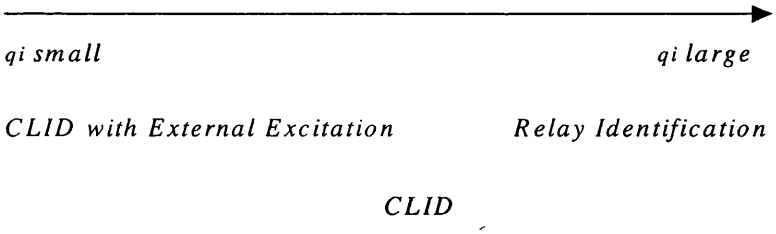


Figure 4-9 *CLID methods suitable for different quantization interval (qi)*

Small quantization interval

If the quantization interval is small, the quantizer in Matlab Simulink is sufficient since the hysteresis will not make any difference in signal generation. On this occasion, the set-point of the closed-loop must be one of the quantization levels (*ql*) and the hysteresis ε should be equal to zero.

For a given process, there are conflicts in choosing the quantization interval when the quantization interval is small (i.e. for the left side of Figure 4-9). If the quantization interval is small, the quantization error excitation will be white (therefore persistently exciting) and uncorrelated with the process disturbance, but the signal-to-noise ratio is not large enough to do CLID (just like the spectral power as indicated in Figure 4-7). Otherwise, if the quantization interval becomes larger, the quantization error excitation will be large enough, but the correlation between the quantization error excitation and the process disturbance will be larger. This will also endanger the CLID. Therefore, it is a challenge to determine the appropriate range of quantization intervals.

Large quantization interval

If quantization interval is large (i.e. for the right side of Figure 4-9), the main consideration is to guarantee the quantizer behaves the same as a relay since one of the relay-based identification methods will be used. On this occasion, none of the quantization levels (*ql*) can be equal to the setpoint of the closed-loop since a mid-rise quantizer will be used. The hysteresis is also necessary because it is the main means to prevent disturbance influence. The hysteresis should be determined according to the size of the disturbance. This will be discussed in more detail in subsection 4.3.3.

4.3.2 Closed-loop identifiability for different quantization intervals

The purpose of this subsection is to indicate that different part of quantization interval (as shown in Figure 4-9) corresponds to different reasons for the ability to do CLID.

Closed-loop identifiability has been reviewed in Chapter 2. Cases relating to closed loop identifiability have been studied in Chapter 3. A fundamental problem with closed-loop data is the correlation between the unmeasurable noise and the process input. It is necessary to break the dependency between the process input and the process noise. From the discussions in Chapter 2, closed-loop identification can be accomplished in one of the following three ways:

- By injecting an independent, persistently exciting signal into the feedback loop;
- By switching between two or more feedback controllers;
- By inserting nonlinearity.

Small quantization interval

For the left part in Figure 4-9, when the quantization interval is small, the quantization error is persistently exciting. Thus, inserting a quantizer will have the same effect as injecting an independent, persistently exciting signal.

Large quantization interval

For the right part in Figure 4-9, when the quantization interval is large, the quantizer acts like a relay. At this moment, the nonlinearity will be dominant. So, inserting a quantizer will have the same effect as inserting a nonlinearity.

4.3.3 Key issues for CLID with a quantizer

Small quantization interval

As discussed in Subsection 4.3.1, when the quantization interval is small, CLID with External Excitation will be used. There is a trade-off between the ***signal-to-noise ratio*** and the ***correlation*** between the signal and noise in determining the appropriate range of quantization intervals. Here, the signal refers to the quantization error and the noise refers to the process disturbance. The task is to find a value or a range of the values for the quantization interval (*qi*), which will make the quantization error good enough for CLID with External Excitation (e.g. the two-step or two-stage method as discussed in Chapter 3) to be used.

Simulation in Chapters 5 and 6 (Figure 4-10 is the same as Figure 6-2) shows a typical relationship of the *signal-to-noise ratio* and the *correlation* as a function of the quantization interval.

From the simulation result shown in Figure 4-10, the useful range of quantization interval should be from σ_y to $2\sigma_y$ approximately, in which σ_y denotes the standard deviation of the process output. The simulation suggests that when the quantization interval is within this range, the identified model accuracy can be guaranteed.

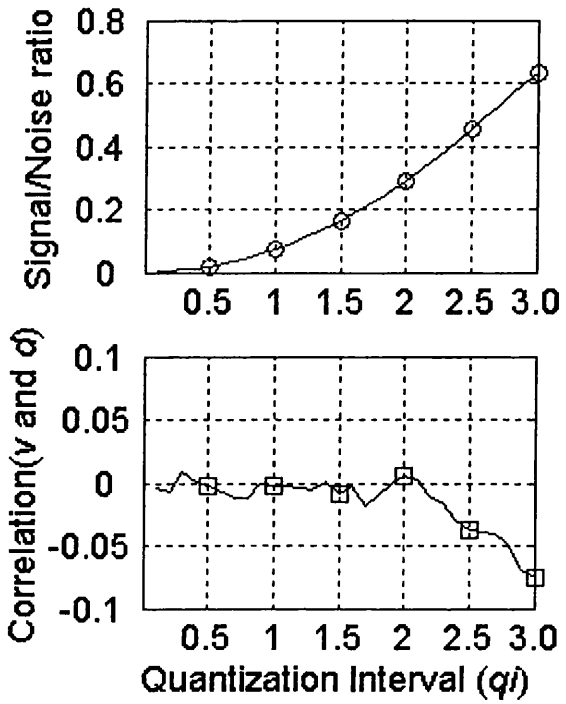


Figure 4-10 *The relationship of the signal-to-noise ratio with qi for a given process and the relationship of the correlation of the quantizer error excitation and disturbance with quantization interval (qi).*

Figure 4-11 which is the same as 6-3 is generated from the simulation scheme in Figure 6-1 in which the disturbance is white noise with variance 1.00 and the quantization interval is 1.5 times the standard deviation of the process output ($qi = 1.5\sigma_y$). Figure 4-11 is a typical example of an unquantized signal, quantized signal and quantization error in the time domain. In the same figure, the quantization error is expressed in the frequency domain. From the left bottom panel of the figure, it can be seen that the quantization error is random. From the right bottom panel, the quantization error is persistently exciting.

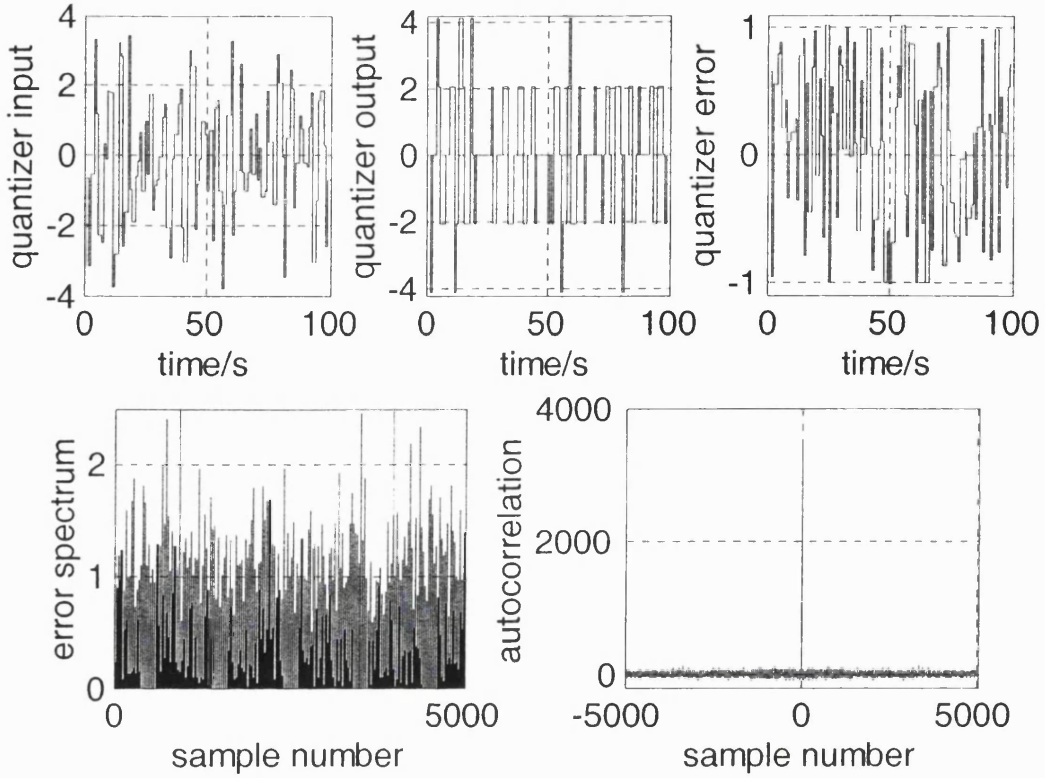


Figure 4-11 *Unquantized signal (left top panel), quantized sigal (middle top panel) and quantization error (right top panel), all three panels in the top are in the time domain; the power spectrum of the quantization error (left bottom panel), autocorrelation of the quantization error (right bottom panel), all two panels are in the frequency domain.*

Large quantization interval

The discussion followed in this section shows that filtering is preferable to hysteresis since filtering reduces the requirement of the quantization interval for CLID when a measurement noise exists.

As discussed in subsection 4.3.1, when the quantization interval (qi) is large, the quantizer will act like a relay and the main consideration is to guarantee that the ***quantizer behaves the same as a relay***. Therefore, only two quantization levels can be used.

Noise is inevitable in a control loop. The general practice is to use hysteresis as the main measure to prevent disturbance influence. The hysteresis width is usually chosen as two times larger than the noise magnitude (Wang *et al*, 1997).

The analysis in Welsh and Goodwin (1999) based on Atherton (1975) is very helpful. They show that adding hysteresis moves the describing function of the relay down the imaginary

axis (the left panel of Figure 4-12). Therefore, the intersection of the describing function and a plant transfer function will generally occur at a frequency less than the critical frequency.

As discussed in subsection 3.3.1, the describing function $N(a)$ of the ideal relay with relay amplitude d and amplitude of limit cycle a is given by

$$N(a) = \frac{4d}{\pi a} \tag{4.1}$$

If hysteresis is used with hysteresis width (ε) for the ideal relay, the describing function $N(a)$ becomes (Atherton, 1975)

$$N(a) = \frac{4d}{\pi a^2}(\sqrt{a^2 - \varepsilon^2} - j\varepsilon) \text{ for } a \geq \varepsilon \tag{4.2}$$

$$-\frac{1}{N(a)} = -\frac{\pi a}{4d}(\sqrt{1-(\frac{\varepsilon}{a})^2} + j\frac{\varepsilon}{a}) \tag{4.3}$$

For the relay-based identification method to work, the system must satisfy

$$G(j\omega_u)N(a)+1=0 \tag{4.4}$$

That is to say, the system frequency response $G(j\omega)$ and minus the inverse of the describing function $-1/N(a)$ should have an intersection. From Figure 4-12, the inverse of the describing function $-1/N(a)$ is a locus in the Nyquist plane parallel to the negative real axis. The position of the locus with respect to the imaginary axis is a function of the hysteresis width and the amplitude of the limit cycle.

From the left panel of Figure 4-12, for a given relay amplitude, when the hysteresis width is larger than a specific value, there is a possibility that no intersection exists between minus the inverse of the describing function and the plant transfer function. This will make relay-based identification impossible. The way to increase the likelihood of having intersection between minus the inverse of the describing function and the plant transfer function is to increase the relay amplitude as shown in the right panel of Figure 4-12. Therefore, relay

amplitude and hesteresis width should maintain a minimum ratio for the relay-based identification.

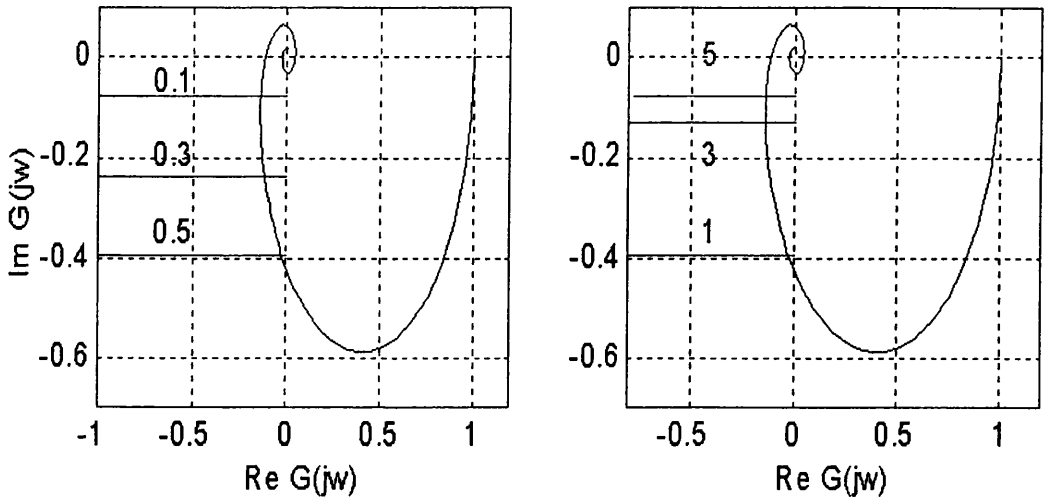


Figure 4-12 *Influence of the hysteresis width and relay amplitude in relay-based identification. Left: the relay amplitude is fixed to be 1, the hysteresis width is changed from 0.1 to 0.3, then to 0.5. Right: the hysteresis width is fixed to be 0.5, the relay amplitude is changed from 1 to 3 and then to 5.*

For a white measurement noise (for example, a in Figure 4-8) with standard deviation σ, according to the normal distribution theory shown in Figure 4-13 (Scheaffer and McClave, 1995), there is 99.7% chance for noise value to be within $\pm 3\sigma$ at each instant. Therefore, 6σ as the hysteresis width should be a reliable choice.

For the white measurement noise (a in Figure 4-8) with standard deviation σ, if the hysteresis width (6σ) is used, then the relay amplitude required will be at least 18σ in order to have an intersection between minus the inverse of the describing function and a plant transfer function.

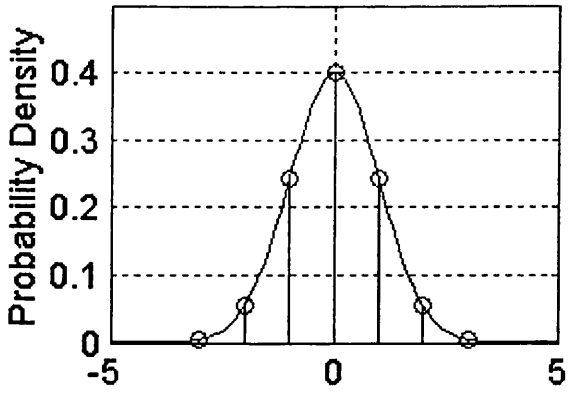


Figure 4-13 *Normal distribution theory; here a standard normal distribution with standard deviation $\sigma = 1$ and the mean value $u = 0$; the area (i.e. the integrated probability density) between [-1, 1] is 68%, [-2,2] is 95%, [-3,3] is 99.7%.*

From Åström and Hagglünd (1984**a**), filtering is another means to prevent noise influence. As discussed in subsection 3.3.3 of Chapter 3, with a lowpass Butterworth filter, most of the noise will be filtered. The hysteresis will only be a secondary measure to prevent noise influence and the hysteresis width required is much smaller. The new scheme with an online filter is shown in Figure 4-14. The advantage of the lowpass Butterworth filter to prevent measurement noise is that the hysteresis width will be smaller and the quantization interval required could also be smaller since the analysis of Figure 4-12 indicated that the relay amplitude and hesteresis width should maintain a minimum ratio for the relay-based identification.

The low-pass filter can deal with noise above the cutoff frequency (in the frequency domain) effectively. For noise beyond this region, it still relies on the exciting signal power to reduce the noise influence. The relay height is proportional to the signal energy injected into the controlled loop. The higher the relay height, the more the signal power.

When both of the factors have been considered, with this new scheme (Figure 4-14), it is possible to say that any quantization interval above $4\sigma_y$ can be used where σ_y means standard deviation of the process output. The fundamental rules are that the noise power above the filter cutoff frequency will be filtered and the noise power below the filter cutoff frequency will be dealt with by the excitation from the relay (or the quantizer).

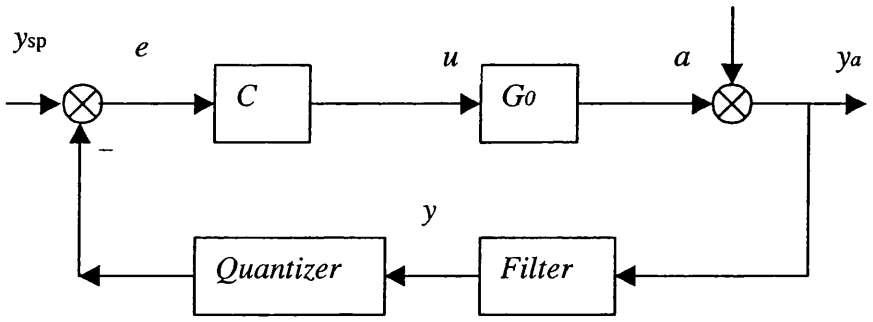


Figure 4-14 *Scheme of CLID with a quantizer*

4.3.4 Nonlinearity

Simulation examples on quantized signals in subsection 4.2.3 indicate that the nonlinearity index becomes larger with the increase of the quantization interval. Simulation examples presented in subsection 6.2.3 below also support this conclusion. One way to break the dependency between the input signal and the process disturbance in a closed loop is to insert nonlinearity (Ljung, 1987). Therefore, nonlinearity caused by quantizer is helpful for closed loop identification.

4.3.5 Further recommendations on quantization interval selection

The purpose of this subsection is to refine the range of the quantization interval when it is small (i.e. for the left side of Figure 4-9). The suggestion in subsection 4.3.3 is that the useful range of the quantization interval should be from σ_y to $2\sigma_y$ where σ_y denotes the standard deviation of the process output. Since the process noise, the noise model and the controller all contribute to the process output, further recommendations in this section are based on the observation that the minimum variance for the process output (σ_{MV}) is a better reflection of the noise applied to the process than the standard deviation of the process output σ_y. Therefore, the standard deviation of the minimum variance for the process output (σ_{MV}) should be the target.

According to the closed-loop performance assessment theory presented in section 2.4 of Chapter 2, σ_{MV} can be calculated with routine process data (i.e. without any excitation). In the new procedure of CLID with a quantizer (as described in subsection 5.1.3), the first step

is to find the standard deviation of the process output σ_y. Then the same data set is used to calculate the standard deviation of the minimum variance for the process output (σ_{MV}).

The procedures of the proposed method are presented in subsection 5.1.3. The simulation examples with three different controllers for the same process and the corresponding results are presented in subsection 5.4.3 and subsection 6.1.2. These examples validated that σ_{MV} is a better reflection of the process than σ_y. The further recommendation is that the quantization interval should be from $1.15\sigma_{MV}$ to $1.95\sigma_{MV}$ when the quantization interval is small.

4.3.6 Selection of the identification methods

It is always useful to use the suggestions proposed in subsection 4.3.5. However, the motivation of this thesis is to find a novel way to do CLID with as little disruption to the process as possible. As presented in section 6.5, the insights from simulation examples and experiment are that: In order to minimize the disruption to the process, when the noise is small, relay-based identification is preferred; when the noise is large, closed-loop identification with external excitation method is recommended.

In real practice, the noise itself cannot be measured or predicted. Its influence will be reflected in the measurement of the process output. The way to judge small or large noise is to see the variance of the process output. The criterion to judge small or large noise is to compare the value of the variance of the process output being that small noise refers to σ_y^2 less than 0.1 on the 4 – 20 mA scale and large noise refers to σ_y^2 greater than 0.1 on the 4 – 20 mA scale (y is the process output as shown in Figure 4-8 in p125).

4.3.7 Summary

The suggestion in subsection 4.3.3 for closed loop identification with external excitation when the quantization interval is small is that the useful range of quantization interval should be from σ_y to $2\sigma_y$. In subsection 4.3.5, the further suggestion for closed loop identification with external excitation when that quantization interval is small is that the useful range of quantization interval should be from $1.15\sigma_{MV}$ to $1.95\sigma_{MV}$. The further

recommendation is based on the observation from the simulation examples that σ_{MV} is a better measure for specifying the quantization interval for CLID than σ_y.

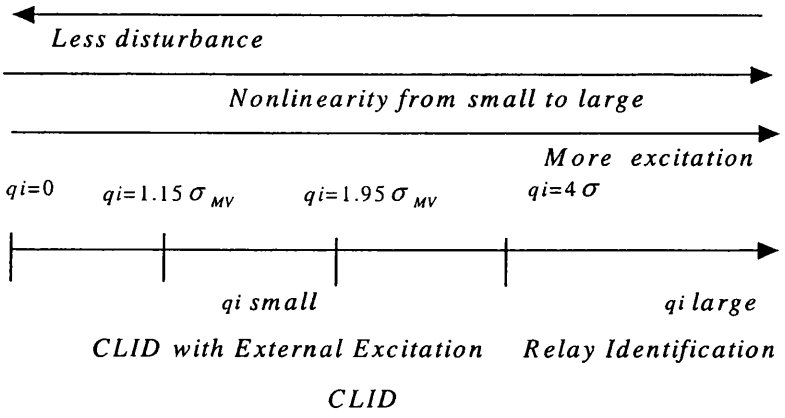


Figure 4-15 *CLID methods suitable for different quantization interval (qi)*

The conclusion in subsection 4.3.3 for relay-based identification when the quantization interval is large is that the useful range of quantization interval should be any value above $4\sigma_y$.

4.4 Effect on Stability of the newly inserted quantizer

The purpose of this subsection is to analyze the effect on closed-loop system stability by the newly inserted quantizer in the scheme for closed-loop identification with a quantizer as shown in Figure 4-8.

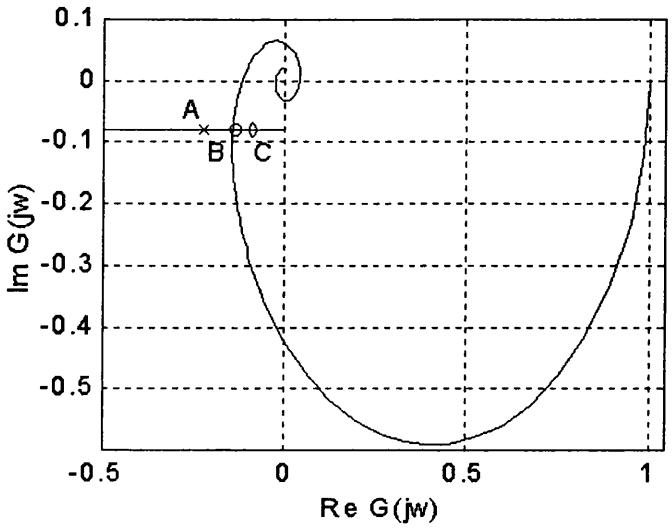


Figure 4-16 *Stability analysis after inserting a relay with hysteresis. 'B' is the point with a circle where the minus inverse of the describing function intersects with*

the process Nyquist plot; 'A' is the point with a x-mark; 'C' is the point with a diamond.

When the quantization interval is small, the quantizer used is a mid-step quantizer, the quantizer can be equivalent as a mechanism to generate the excitation signal.

When the quantization interval is large, the quantizer used is a mid-rise quantizer, the stable state of the system is a limit cycle (Atherton, 1975) (point 'B' in Figure 4-16). If the closed-loop system deviates from this state, for example, if the limit cycle amplitude becomes smaller due to some disturbance, point 'C' in Figure 4-16, the closed-loop system will become unstable. This will make the limit cycle amplitude larger. This means returning to point 'B'. On the other hand, if the closed-loop system deviates from this state, for example, if the limit cycle amplitude becomes larger due to some disturbance (point 'A' in Figure 4-16), the closed-loop system will become damped instead of oscillating. This will make the limit cycle amplitude smaller. This means returning to point 'B'.

4.5 Conclusions

Based on the study of the quantizers, the characteristics of the quantization error, a new theory of closed-loop identification based on quantization is proposed. The conclusions are:

- When the quantization interval is small, the closed-loop identification with external excitation can be used; when the quantization interval is large, the relay-based identification can be used.
- In the context of closed-loop identifiability, when the quantization interval is small, the quantization error is equivalent to a persistently exciting external signal; when the quantization interval is large, nonlinearity is dominant in the quantization error.
- When the quantization interval varies from small to large, the power of the quantization error will vary from small to large, while the bandwidth of the quantization error will vary from wide to narrow.
- When the quantization interval varies from small to large, the nonlinearity of the quantization will be from small to large.

CHAPTER 5

METHODS

In this chapter, the two-stage method (Van den Hof and Schrama, 1993) and the two-step method (Huang and Shah, 1997), which were presented in Section 3.2, are first adapted for CLID with a quantizer. Then, one of the relay-based identification methods, which was proposed by Wang *et al.* (1997b) as presented in Section 3.3, is also adapted for the new identification purpose. Other methods, including model accuracy measure and nonlinearity test, are discussed. Simulation methods include transfer function model simulation and pilot plant model simulation, which will be detailed in Section 5.4, Section 5.5 and Section 5.6 respectively. Experimental validation with a pilot plant is presented in Section 5.7. A summary is given at the end. This chapter presents the methods. All results will be presented in Chapter 6.

5.1 Closed-loop identification with external excitation

5.1.1 Adaptation of the closed-loop identification with external excitation method

The two-stage method (Van den Hof and Schrama, 1993) and the two-step method (Huang and Shah, 1997) were presented in Chapter 3. The dither signal is injected after the controller and before the process (as shown in Figure 5-1).

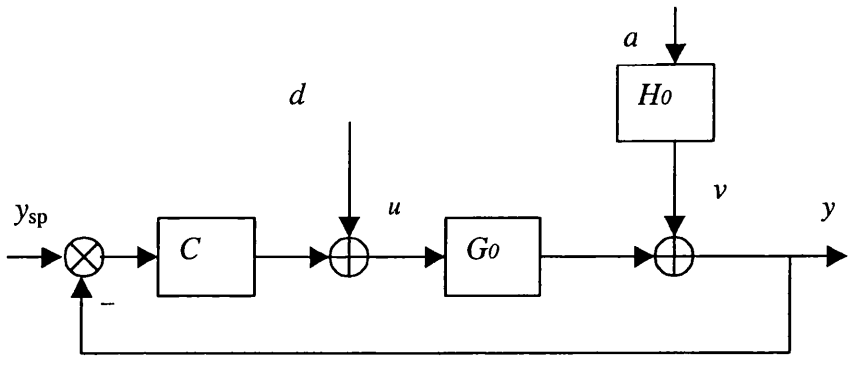


Figure 5-1 *Closed-loop system with excitation in the forthward path*

The system with a dither signal in the feedback path as shown in Figure 5-2 is considered in this chapter and the following chapter. Therefore the task is to adapt these two methods for further study (i.e. a dither signal in the feedback path).

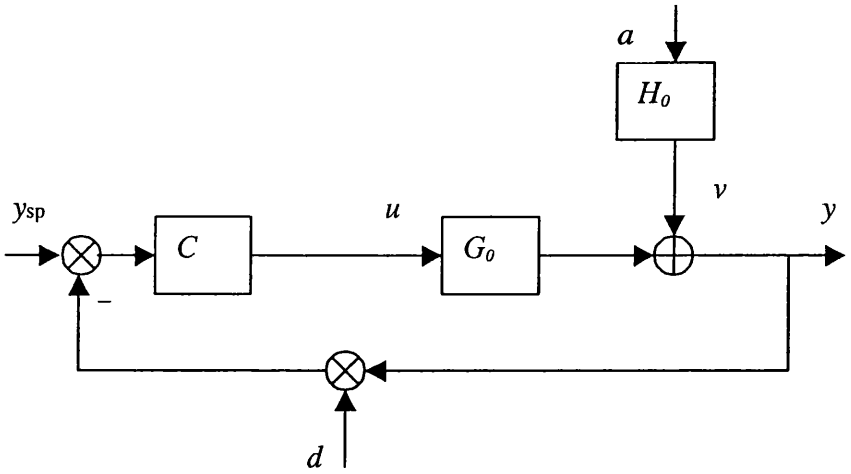


Figure 5-2 *Closed-loop system with excitation in the feedback path*

$$u(t) = \frac{-C(q)}{1+C(q)G_0(q)}d(t) - \frac{C(q)}{1+C(q)G_0(q)}v(t) \qquad (5.1)$$

$$y(t) = \frac{-C(q)G_0(q)}{1+C(q)G_0(q)}d(t) + \frac{1}{1+C(q)G_0(q)}v(t) \qquad (5.2)$$

From (5.1) and (5.2)

$$u(t) = -C(q)S_0(q)d(t) - C(q)S_0(q)H_0(q)a(t) \qquad (5.3)$$

$$y(t) = -C(q)G_0(q)S_0(q)d(t) + S_0(q)H_0(q)a(t) \qquad (5.4)$$

where S_0 is the true sensitivity function, which equals to $\frac{1}{1+G_0C}$.

Similar to the presentation in Section 3.2, the procedures of the two-stage method by Van den Hof and Schrama (1993) are:

- The product of the sensitivity function and the controller are obtained from the excitation d and the process input u. Then, a noise-free process input u^d is simulated from excitation d and the product of the sensitivity function and the controller;
- The estimate of the process G_0 is calculated from the noise-free process input u^d and the process output y:

$$u(t) = [-C(q)S_0(q)]d(t) + [-C(q)S_0(q)]H_0(q)a(t) \qquad (5.5)$$

$$u^d(t) = [-C(q)S_0(q)]d(t) \qquad (5.6)$$

$$y(t) = G_0(q)u^d(t) + S_0(q)H_0(q)a(t) \qquad (5.7)$$

Also from Section 3.2, the procedures of the two-step method by Huang and Shah (1997) are:

- The product of the sensitivity function and the controller are obtained from the excitation d and the process input u ;
- The estimate of the process G_0 is calculated from the process input u and the process output y filtered by the product of the sensitivity function and the controller.

$$u(t) = [-C(q)S_0(q)]d(t) + [-C(q)S_0(q)]H_0(q)a(t) \qquad (5.8)$$

$$\frac{y(t)}{-C(q)S_0(q)} = G_0(q)d(t) + \frac{H_0(q)a(t)}{-C(q)} \qquad (5.9)$$

Remark: When the dither signal position is in the feedback path, the two-step method and the two-stage method have the same procedures, the only differences are that *–C(q)S(q)* instead of *S(q)* is identified from *u(t)* and *d(t)* in the first step and that *–C(q)S(q)* instead of *S(q)* is then used in the second step.

5.1.2 Closed-loop identification with external excitation using a quantizer

A quantizer is inserted in the feedback path (Figure 5-3) to generate the equivalent dither signal whose characteristics were discussed in the theory section (subsection 4.3.3). Therefore the adapted method, as shown in equations (5.5), (5.6) and (5.7) or equations (5.8) and (5.9), can be used directly.

As shown in Figure 5-3, the dither signal d is generated by the quantizer. Therefore, the dither signal d is quantization error excitation which is defined as follows:

$$d = y_q - y \qquad (5.10)$$

where y is the process output, which is the unquantized signal and y_q is the quantized signal.

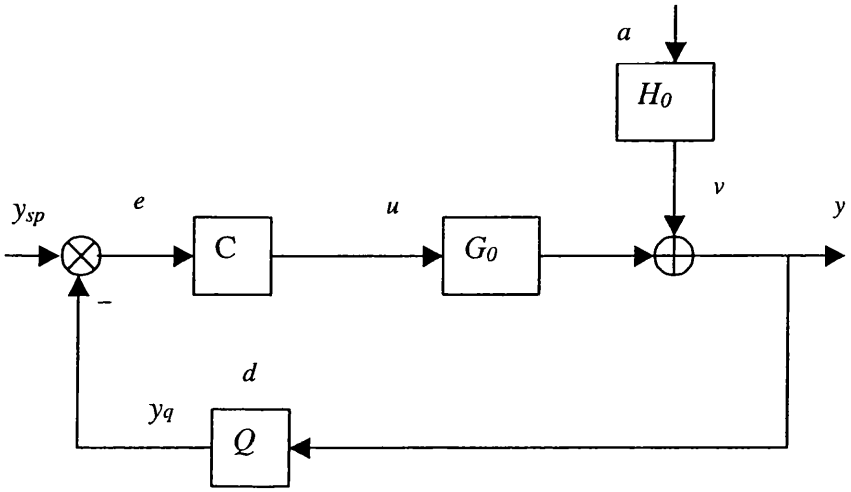


Figure 5-3 *The proposed CLID with a quantizer scheme*

The procedures for the proposed CLID scheme (CLID with a quantizer at the feedback path) were as follows:

- The process was run as normal (with quantization interval very small, for example 0.005, equivalent to 12 bit A/D). The standard deviation of process output σ_y is calculated.
- A specific quantization interval (in the range of σ_y~$2\sigma_y$) is chosen. The process is run with this quantization interval. The quantizer error excitation $d = y_q - y$ can be guaranteed to be sufficiently good to do identification, as shown in subsection 5.4.2 and subsection 6.1.1 by simulation.
- The two-step or two-stage method can be used directly.

5.1.3 Further recommendations on quantization interval selection

As analysed in subsection 4.3.5, both the process noise (including the noise model) and the controller contribute to the process output. After considering these two factors, the standard deviation of the minimum variance for the process output (σ_{MV}) should be a useful measure for specifying the quantization interval for CLID. The closed-loop performance assessment (CLPA) theory can be used to calculate σ_{MV}. Simulation examples in subsection 5.4.3 and subsection 6.1.2 indicate that σ_{MV} is a better measure for specifying the quantization interval for CLID than σ_y.

According to Harris (1989) and Desborough and Harris (1992), in Figure 5-3, the controller error e can be expressed as

$$e = \hat{e} + r \tag{5.11}$$

where $\hat{e}$ is forward prediction of controller error and r is the residual.

The closed-loop performance assessment (CLPA) was defined as:

$$\eta(b) = 1 - \frac{\sigma_{MV}^2}{\sigma_e^2} \tag{5.12}$$

where b is the horizon (i.e. how far ahead to do the forward prediction; b is time delay if it is known, otherwise b is an estimate of the time delay), σ_e^2 is the controller error variance and σ_{MV}^2 is the minimum variance achievable.

The criterion to judge good or bad control effect is that σ_e should be close to σ_{MV} so that $\eta(b)$ tends to zero, which indicates good controller performance. Otherwise, $\eta(b)$ tends to one, which means the poor controller performance.

The way to calculate the standard deviation of the minimum variance for the process output (σ_{MV}) with closed-loop performance assessment (CLPA) theory is implemented by Matlab codes (Thornhill *et al.,* 1999 and Desborough and Harris, 1992). The relevant parameters are as follows:

- Determination of prediction horizon *b*.
- Calculation of σ_{MV}, the standard deviation of the minimum variance for the process output, and the CLPA index $\eta(b)$.

The CLID scheme is shown in the Figure 5-3. The procedures for the proposed CLID scheme (CLID with a quantizer) are modified as follows:

- The process was run as normal (with quantization interval very small, for example 0.005mA, equivalent to 12 bit A/D on a 4 – 20 mA scale). The standard deviation of the process output σ_y is calculated.

- The standard deviation of the minimum variance for the process output (σ_{MV}) is calculated with the closed-loop performance assessment (CLPA) theory.
- A mid-step quantizer is used and a specific quantization interval in the range of $1.15\sigma_{MV}$ ~$1.95\sigma_{MV}$ (the illustration is given in subsection 5.4.3 and subsection 6.1.2 by simulation) is chosen. The process is run with this quantization interval. The quantizer error excitation $d = y_q - y$ can be guaranteed to be sufficiently good enough to do identification.
- The two-step or two-stage method can be used directly.

5.2 Relay-based identification method

The FFT-based relay identification method by Wang *et al.* (1997b) is chosen here. The main steps are briefly outlined as follows (For more details, refer to Section 3.3):

- An exponential decay $e^{-\alpha t}$ (α>0) is introduced. Applying the Fourier Transform gives

$$\tilde{Y}(j\omega) = \int_0^\infty y(t)e^{-\alpha t}e^{-j\omega t}dt = Y(j\omega+\alpha) \tag{5.13}$$

$$\tilde{U}(j\omega) = \int_0^\infty u(t)e^{-\alpha t}e^{-j\omega t}dt = U(j\omega+\alpha) \tag{5.14}$$

- The shifted process frequency response $G(j\omega_i + \alpha)$ can be calculated by

$$G(j\omega_i+\alpha) = \frac{Y(j\omega_i+\alpha)}{U(j\omega_i+\alpha)} = \frac{\tilde{Y}(j\omega_i)}{\tilde{U}(j\omega_i)} \tag{5.15}$$

whose inverse FFT is $g(kT)e^{-\alpha kT}$. Thus, the process frequency response $G(j\omega)$ can be calculated.

For the relay identification method by Wang *et al.* (1997b), there is one limitation — the sampling rate has to be fast. The reason is that the samples in one period (or one limit cycle) should be larger than a specific number for FFT-based method to work well. The period is related to the process (for example, the dominant time constant, the time delay) itself. That is why in the case study of Wang *et al.* (1997b) presented in section 3.3, a sampling rate of 0.05 second was used for the process with a dominant time constant of 10 seconds and a time delay of 2 seconds.

5.2.1 Adaptation of the relay-based identification method

In Wang *et al.* (1997b), the relay output is the process input because the controller is switched off. The configuration in this work needs to be different. The relay is replaced by a mid-rise quantizer. A lowpass Butterworth filter is inserted to reduce the influence of measurement noise (more details in subsection 3.3.3 of Chapter 3). As indicated in subsection 4.3.3.of Chapter 4, the advantage of using a lowpass Butterworth filter plus the relay with hysteresis to reduce measurement noise over only the relay with hysteresis is that the hysteresis width required could be smaller and the quantization interval required could be smaller accordingly.

With this new scheme (as shown in Figure 5-4), the quantizer output is the controller input and the controller output is the process input. Also, the process output includes measurement noise. Only the filter output is useful. The solution to this relay-based identification problem is to view the process, the controller and the filter as a whole new process. In other words, y_q is viewed as input and y_f is viewed as output for relay-based identification with Wang *et al.* (1997b). Therefore, it is necessary to separate the controller dynamics and the filter dynamics from the result identified.

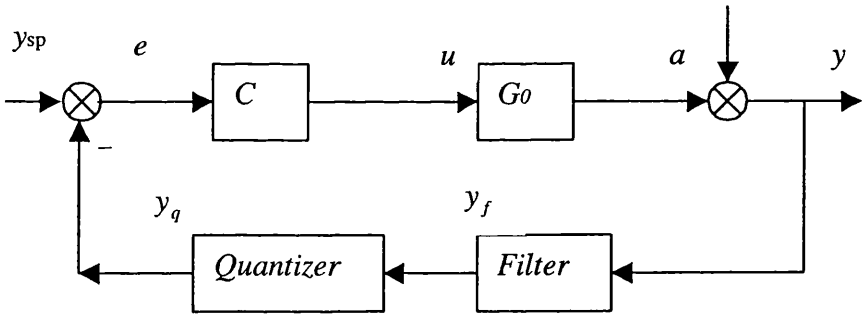


Figure 5-4 *Scheme of relay-based identification with a quantizer*

5.2.2 Relay-based identification with a quantizer

The procedure for the proposed CLID scheme based on Wang *et al.* (1997b) is as following:

- The process was run as normal (with quantization interval very small, for example 0.005, equivalent to 12 bit A/D). The standard deviation of process output σ_y is calculated.
- A mid-rise quantizer is used and a specific quantization interval (in the range of greater than $4\sigma_y$) is chosen together with the hysteresis width and paramteters for the lowpass

Butterworth filter. The process is run with this quantization interval. The quantizer output and the filter output are logged.

- The relay-based identification method (Wang *et al.*, 1997b) is used directly.
- The controller dynamics and the filter dynamics are separated (by complex dividing) from the result identified from the third step in order to get the process estimate.

5.3 Computation methods

5.3.1 Model accuracy measure

For assessment of the accuracy of the estimate, the identification error in this thesis is measured by the worst-case error.

$$ERR\% = \left| \frac{\hat{G}(j\omega_i) - G(j\omega_i)}{G(j\omega_i)} \right| \times 100\% \quad i = 1,2,\cdots M \tag{5.13}$$

In (5.13), $G(j\omega_i)$ and $\hat{G}(j\omega_i)$ are the actual and the estimated process frequency response respectively. Only the part of the Nyquist curve from zero frequency to the crossover frequency ω_{cg} where the argument is –180° was considered since this part is the most important for system identification and controller design.

5.3.2 Nonlinearity test method

Principle for nonlinearity test

The principle of non-linearity test is based on the following three main elements: Null hypothesis, surrogate data and nonlinear prediction (Kantz and Schreiber, 1995).

An appropriate null hypothesis for many nonlinear dynamics tests is that the time trend arises from a linear dynamic system. In order to establish the significance of a test against this null, one can generate many realizations of the Null, and estimate the significance.

The Null of 'linear dynamic system' is not specific. One way to make a specific Null is to set the mean and variance to the same values as those of the original data. Surrogate data is random data generated to have the same mean, variance, and autocorrelation function as the original data. In other words, surrogate data has the same power spectrum as the original data, but the phase relationships that arise from nonlinearity are randomised.

Matlab codes to test nonlinearity

The above principle is implemented with Matlab codes provided by the author's supervisor (Thornhill, N.F., Shah, S.L., and Huang, B., 2001). In the algorithm, surrogate data sets that have the same power spectrum and auto-covariance as the test data set are created first. A suitable sub-set of the data is found with a good end-to-end match. Then the surrogates are calculated. 50 surrogate data sets are used for the test in this thesis. The non-linear zero prediction error of the surrogates is compared with the prediction error for the test data. If the test data were generated by a non-linear process, they are generally more predictable than the surrogates and the error is significantly smaller. "Significant" means lying more than two standard deviations from the mean error of the surrogates.

The result of the algorithm gives the nonlinearity index. It is the prediction error of the test data minus the mean error of the surrogates divided by 2σ. If the data are much more predictable than the surrogates then the value of nonlinearity index is bigger than 1. If the data are roughly as predictable as the surrogates then the value of the nonlinearity index is typically less than 1.

5.4 Discrete transfer function model simulation method

In order to demonstrate the theoretical analysis and the methods discussed, a second order ARMAX model presented in Huang and Shah (1997) was chosen as a simulation example. The purpose of this section is to emphasize the validation of the left side of Figure 4-9 (or Figure 4-15, i.e. when the quantization interval is small) and to emphasize the procedures for further recommendations on the selection of the quantization interval.

5.4.1 Simulation example

A second order ARMAX model presented in (Huang and Shah, 1997) was chosen to validate the theoretical analysis and the methods discussed. The transfer function was:

$$y(t) = \frac{0.3403 + 0.2417q^{-1}}{1 - 0.7859q^{-1} + 0.3679q^{-2}}u(t-1) + \frac{1 - 0.8q^{-1} + 0.12q^{-2}}{1 - 0.7859q^{-1} + 0.3679q^{-2}}a(t) \quad (5.17)$$

A unit feedback control law is implemented in this simulation (Figure 5-5). The white noise $a(t)$ is applied. The number of sampled data points in the simulation was 10,000.

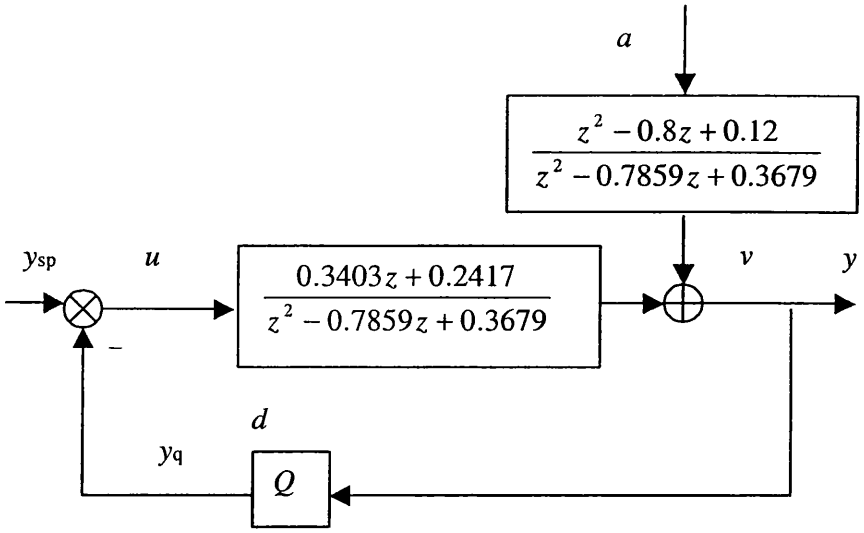


Figure 5-5 *Simulation for CLID with a quantizer for discrete transfer function model*

5.4.2 Correlation and signal-to-noise ratio test

Simulation in this subsection with Figure 5-5 is used for the purpose of exploring the characteristics of the quantizer excitation under different quantization intervals. In this simulation, a white noise with σ_a^2=1 is applied and fixed throughout the tests. The standard deviation of the process output is σ_y=1.373 when the quantization interval is 0.005 (the same as Case 3 of the following subsection 5.4.4). There are altogether 30 tests ranging from *qi*=0.10, *qi*=0.20, *qi*=0.30 until *qi*=3.00. The results will be presented in subsection 6.1.1.

5.4.3 Details for further recommendations on quantization interval selection

The purpose of this subsection is to illustrate the theory proposed in subsection 4.3.5 of Chapter 4. That is to explain the details on further recommendations on quantization interval selection in subsection 5.1.3, i.e. the quantization interval should be from $1.15\sigma_{MV}$ to $1.95\sigma_{MV}$ when the quantization interval is small. Simulations with Figure 5-4 are conducted for different proportional-only controllers (Table 5-1). Therefore all the simulation parameters are the same as subsection 5.4.2 except the controller. The results will be presented in subsection 6.1.2.

	Case 1	Case 2	Case 3
Controller Gain	1.00	0.50	2.00

Table 5.1 *CLID with a quantizer for different controllers*

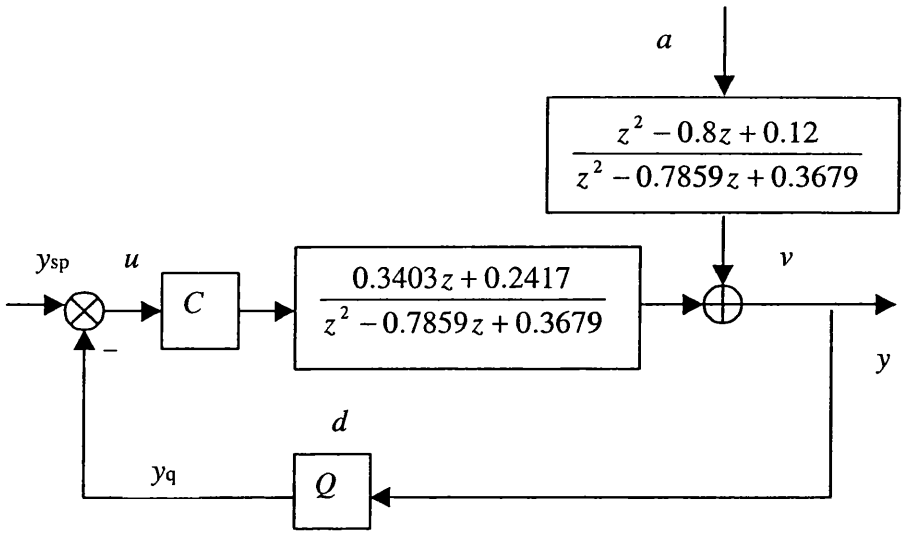


Figure 5-6 *Simulation for CLID with a quantizer for discrete transfer function model (with different P-only controllers)*

5.4.4 CLID with external excitation using a quantizer

The purpose of this subsection is to explore the proposed CLID with a quantizer method under different disturbances and different quantizer intervals. The following three cases (Table 5.2) were designed and tested with simulation. Therefore, the influence of the disturbance size and quantizer interval can be observed from the results. The results will be presented in Subsection 6.1.3

	Case 1	Case 2	Case 3
Variance of disturbance $a(t)$: σ_a^2	0.010	0.010	1.000
Disturbance equivalent to 4-20mA	0.100	0.100	1.000
Standard deviation of process output when qi = 0.005mA	0.137	0.137	1.373
Quantization interval (qi) for identification	0.137	0.2025	1.373
Ratio of qi for identification with standard deviation	1	1.5	1
Quantization interval (qi) equivalent to A/D bit	7	6	3 or 4

Table 5.2 *CLID with a quantizer simulation conditions for discrete transfer function model simulation*

5.5 Continuous transfer function model simulation method

The purpose of this section is to emphasise the illustration of the quantization interval range proposed in Figure 4-9 as well as in Figure 4-15 (i.e. for a given process and a given disturbance, when the quantization interval is small, closed loop identification with external excitation method can be used; when the quantization interval is large, relay-based identification can be used) and to indicate nonlinearity from small to large for the quantization intervals from small to large.

5.5.1 Simulation example

The commonly used continuous-time transfer function model (5.18) was chosen (e.g., Li *et al.* 1991, Wang *et al.* 1997) to demonstrate the theoretical analysis and the methods discussed above. This model has been widely used in Chapter 3 of this thesis.

$$G_{process} = \frac{e^{-2s}}{10s+1} \tag{5.18}$$

5.5.2 CLID with external excitation using a quantizer

This subsection demonstrates closed-loop identification with external excitation using a quantizer method in the process for the continuous transfer function (5.18). The simulation scheme is shown in Figure 5-7.

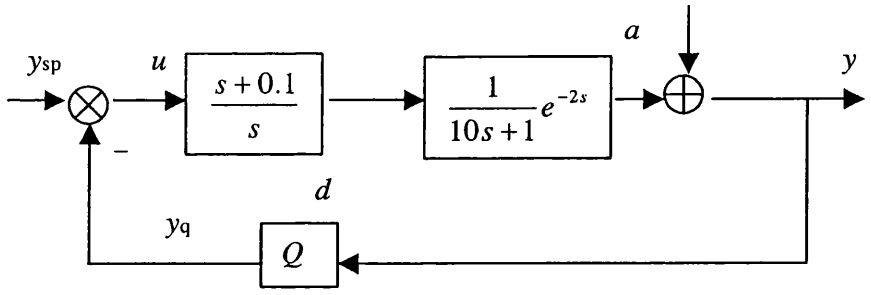


Figure 5-7 *Simulation for CLID with a quantizer for continuous transfer function*

In Figure 5-7, the disturbance a is a white noise with variance 0.10. The quantization error $d = y_q - y$ is the exciting signal generated by the quantizer, which is persistently exciting. The quantizer works as a mid-step quantizer. The two-stage method (Van den Hof and

Schrama, 1993) was used for identification. The results will be presented in Subsection 6.2.1.

5.5.3 Relay-based identification using a quantizer

The simulation in this subsection is to demonstrate the relay-based identification using a quantizer for the continuous transfer function model. The identification scheme is shown in Figure 5-8.

In the block diagram, a is a white noise signal with variance 0.10, which is implemented by the Band-limited White Noise module in Matlab/Simulink. At this moment, the quantizer works as a mid-rise quantizer. Only two quantization levels can be used during the identification. The FFT-based method (Wang *et al*, 1997b) was used for identification. The quantizer output y_q (instead of the process input u) is viewed as input. The output of the low-pass Butterworth filter y_f (instead of the process input y) is viewed as output. The dynamics of the controller and the filter are removed from the identified results.

The standard deviation of the process output is $\sigma_y = 0.32$ when the quantization interval is 0.005mA on the 4 – 20 mA scale. According to the CLPA theory, the standard deviation of the minimum variance for the process output is $\sigma_{MV} = 0.32$. By the suggestions in Chapter 4, a quantization interval equal to $4\sigma_y$ is used. The results will be presented in subsection 6.2.2.

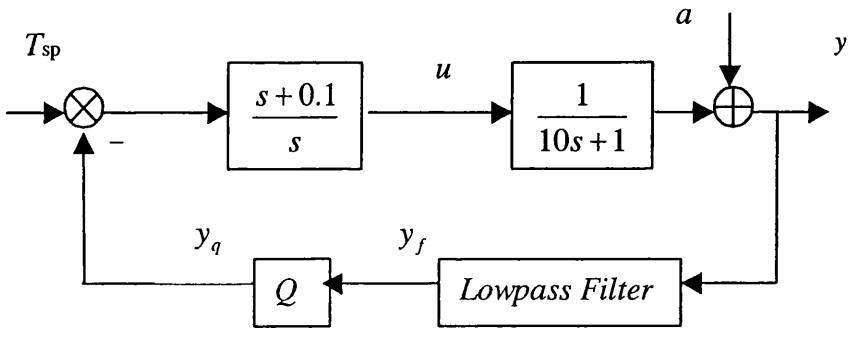


Figure 5-8 *Relay-based identification with a quantizer for the continuous transfer function model.*

5.5.4 Nonlinearity test

The nonlinearity test was described in subsection 5.3.2. The purpose of this subsection is to show the nonlinearity index changes with the two simulation examples presented in subsection 5.5.2 and 5.5.3 in which quantization interval is 0.48 and 1.28 respectively (i.e. $1.5\sigma_y$ and $4\sigma_y$ respectively). The results will be presented in Subsection 6.2.3.

5.6 Pilot plant model simulation method

The purpose of this section is to arrange simulation examples for validating the quantization interval range proposed in Figure 4-9 (i.e. for a given process and a given disturbance, when the quantization interval is small, closed loop identification with external excitation method can be used; when the quantization interval is large, relay-based identification can be used) with a real life simulation model instead of the transfer functions used before and to prepare for the next section. All the measurements in this section are on the 4 – 20 mA scale except those explicitly given.

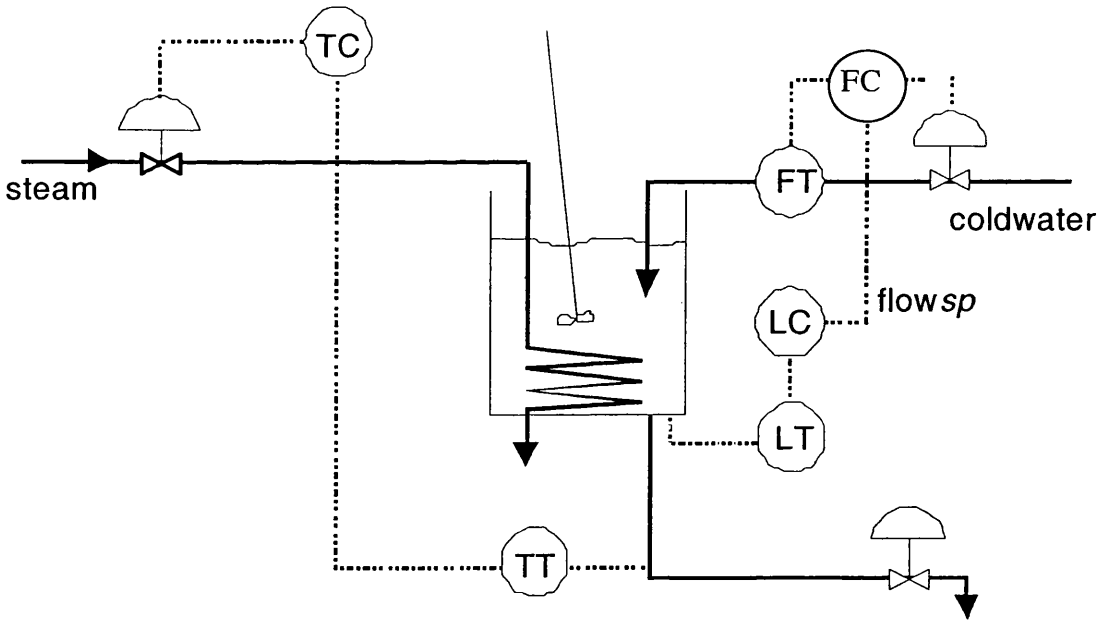


Figure 5-9 *The experimental apparatus*

5.6.1 Introduction of the pilot plant and its simulation model

An experimental apparatus (as shown in Figure 5-9) was used to evaluate the proposed scheme. It is a doubled-walled glass tank. The tank has a height of 50 cm and an inside diameter of 14.5 cm. This system is located in the University of Alberta computer process control laboratory.

The author's supervisor provided this simulation model, which can be used in Matlab/Simulink. The examples of closed-loop identification with a quantizer in this section are based on this Simulink model.

5.6.2 Open loop test

The aim of this subsection is to give a benchmark for assesment of CLID results since the true process of the experimental apparatus is not known.

For the open loop test of the temperature loop, two simulation examples were designed. The test signal was applied at the steam valve and the output is measured from the temperature sensor in the outlet pipe of the tank. In the first test, the test signal is a random binary signal (RBS) with frequency band from 0 to 1 (i.e. 0 to 1.57 rad/sec since the sampling time used is 4 second) and the amplitude from −1mA to 1mA. The RBS signal is generated by a function in the System Identification Toolbox in Matlab. In the second test, the input is a unit amplitude step from 12.57mA to 13.57mA. The open loop identification results are shown in subsection 6.3.1.

5.6.3 Closed-loop identification with external excitation using a quantizer

For the demonstration of closed-loop identification with external excitation using a quantizer in the temperature loop of the simulation model, the scheme is shown in Figure 5-10.

In the block diagram, a is white noise with variance 0.25. Quantization error $d = y_q - y$ is the exciting signal generated by the quantizer, which is persistently exciting. At this moment, the quantizer works as a mid-step quantizer. The two-stage method (Van den Hof and Schrama, 1993) was used for identification. The first stage is to identify the sensitivity function with exciting signal d and process input u. In the second stage, the process is identified with simulated input u^d and process output y where simulated input u^d is obtained from exciting signal d and the estimated sensitivity function.

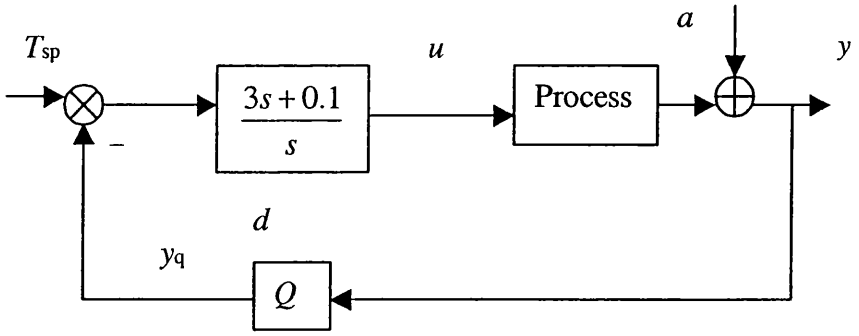


Figure 5-10 *Simulation for the CLID with a quantizer for temperature loop with Van den Hof and Schrama (1993) method.*

The standard deviation of the process output is σ_y=0.55 when the quantization interval is 0.005. The standard deviation of the process output in minimum variance σ_{MV} is 0.52 by calculation. Three cases were designed in which the quantization intervals are different (as shown in Table 5.3). The identification results are presented in subsection 6.3.2.

	Case 1	Case 2	Case 3
Quantization interval (*qi*)	0.61	0.81	1.05
Equivalent to σ_y	1.10	1.47	1.9

Table 5.3 *CLID with a quantizer simulation conditions for pilot plant model simulation*

5.6.4 Relay-based identification with a quantizer

The simulation in this subsection demonstrates the relay-based identification using a quantizer in the temperature loop of the simulation model. The identification scheme is shown in Figure 5-11.

In the block diagram, a is white noise with variance 0.25, which is implemented by the band-limited white noise module in Matlab/Simulink. In this work, the quantizer works as a mid-rise quantizer. Only two quantization levels can be used during the identification. The FFT-based method (Wang *et al*, 1997b) was used for identification. The quantizer output y_q (instead of the process input u) is viewed as the input. The output of the low-pass Butterworth filter y_f (instead of the process input y) is viewed as the output. The dynamics of the controller and the filter are removed from the identified results by complex dividing.

The standard deviation of the process output is $\sigma_y = 0.55$ when the quantization interval is 0.005. According to the CLPA theory, the standard deviation of minimum variance for the process output is $\sigma_{MV} = 0.52$. By the suggestions in Chapter 4, quantization interval equivalent to $4\sigma_y$ is used.

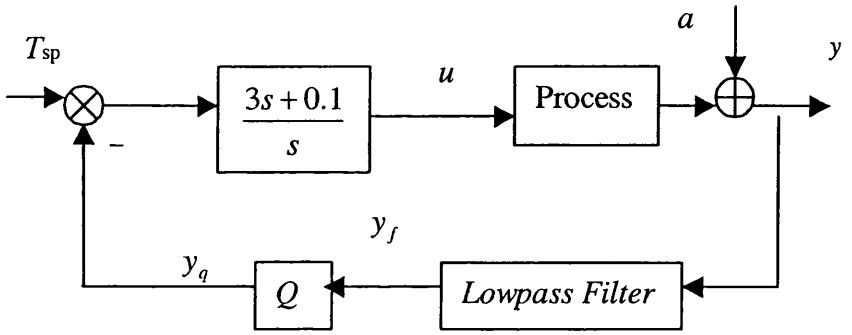


Figure 5-11 *Simulation for relay based identification with a quantization for temperature loop (Wang et al. 1997b).*

5.7 Experimental evaluation

The pilot plant itself has been introduced in subsection 5.6.1. The purpose of this section is to describe the various experiments for demonstrating the method of CLID with a quantizer with a real life pilot plant.

5.7.1 Open loop test

In order to test the dynamics of the temperature loop in pilot plant , two test signals were used respectively, the random binary signal (RBS) and the step. The test signal was applied at the steam valve and the output is measured from the temperature sensor in the outlet pipe of the tank. In the first test, the test signal is RBS with frequency band from 0 to 1 (i.e. 0 to 1.57 rad/sec since the sampling time used is 4 second) and the amplitude from –0.6mA to 0.6mA. The RBS signal is also generated by function in System Identification Toolbox in Matlab. The input is a unit amplitude step from 12.57 to 13.57 in the second test. Two similar step tests have been done. The open loop identification results are shown in subsection 6.4.1.

5.7.2 Closed-loop identification with external excitation using a quantizer

The level of the tank was always controlled to be 12mA on a 4-20 mA scale (that is in the middle of the tank) during the experiment. The cold water inflow and the hot water outflow were always under balance when steady-state was reached. The manipulated variable was the steam flow. The controlled variable was the hot water temperature. A well-tuned PI controller in continuous form was applied (as shown in Figure 5-12, which is the same as Figure 5-10). The setpoint for the temperature loop is 10.5 *mA* (that is 41.5°C).

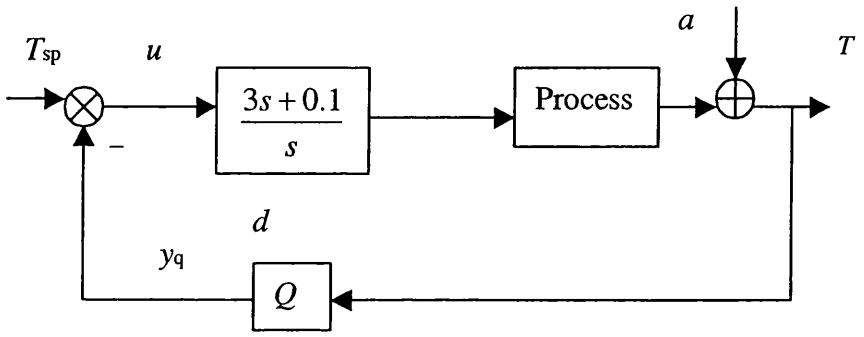


Figure 5-12 *Experimentation for CLID with a quantizer for temperature loop*

Case 1 hybrid experiment

The reason why case 1 is named 'hybrid experiment' is that the process is the real life tank and the computer generated the disturbance a (i.e. by Matlab function). In Figure 5-12, a is white noise with variance 0.25. Quantization error $d = y_q - y$ is viewed as the exciting signal generated by the quantizer. The quantizer works as a mid-step quantizer. The two-stage method (Van den Hof and Schrama, 1993) was used for identification.

The open loop identification result conducted with RBS in subsection 5.7.1 can be viewed as the true process. Then the proposed new scheme is conducted in the following steps:

- The process is run as normal (with quantization interval very small, for example 0.005, equivalent to 12 bit A/D). The standard deviation of process output σ_y is calculated to be 0.55.
- A specific quantization interval qi=1.1 is chosen, i.e. 2 times σ_y. The process is run with this quantization interval. The quantizer error excitation y_q - y can be guaranteed to be sufficiently good to do identification. The signal-to-noise ratio is 0.39.
- The two-step or two-stage method is used directly.

Case 2 natural noise experiment

The difference between this case and Case 1 is the way to generate the disturbance a. In Case 1, the noise is generated with a computer. Here, natural bubbles were applied. Cold air flows through a pipe, which was submerged into the bottom of the tank. This led to observable disturbance to both the level and temperature measurements.

The experimental procedures is as follows:

- The process is run with the bubbles described above (with a very small quantization interval, for example 0.005, equivalent to 12 bit A/D). The standard deviation of the process output σ_y is calculated to be 0.0657.
- A specific quantization interval qi = 0.105 is chosen, i.e. 1.6 times σ_y. The process is run with this quantization interval. The quantizer error excitation y_q - y is used for CLID.
- The two-stage method (Van den Hof and Schrama, 1993) can be used directly.

The identification results for the above two cases are presented in subsection 6.4.2.

5.7.3 Relay-based identification

The relay-based identification method by Wang *et al.* (1997b) needs a sampling rate of at least 0.1 second. The fastest sampling rate in the pilot plant in the University of Alberta is 1 second. Therefore, the method cannot be implemented in experimentation.

5.8 Summary

In this chapter,

- Two typical methods in the closed-loop identification with external excitation framework and the relay-based identification framework are adapted for CLID with a quantizer.
- The model accuracy measure and non-linearity test method has been introduced.
- Transfer function model simulation methods, pilot plant model simulation methods and pilot plant experimental methods have been discussed in detail.

CHAPTER 6

RESULTS AND DISCUSSIONS

All the simulation and experimental results will be presented in this chapter together with discussions. The results and discussions for discrete transfer function examples and continuous transfer function examples are described in section 6.1 and section 6.2 respectively. The results and discussions for simulations with pilot plant model are described in section 6.3. The results and discussions for pilot plant experiments are described in section 6.4. The conclusion will be given in the end.

6.1 Discrete transfer function model simulation

This section is used to describe the simulation results of the discrete transfer function (the second order ARMAX model expressed in 5.1.4). The purpose of the discrete transfer function simulation is to emphasize the left side of Figure 4-9 (i.e. when the quantization interval is small) and to explain the procedures for further recommendations on the selection of the quantization interval. The simulation diagram (Figure 5-5) is represented as Figure 6-1 for convenience of readers.

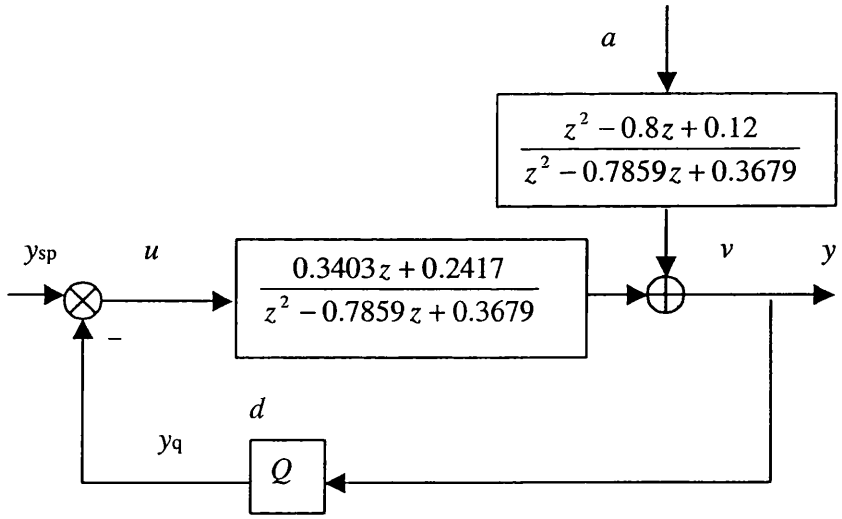


Figure 6-1 *Simulation for CLID with a quantizer*

6.1.1 Correlation and signal-to-noise ratio

The result for the test by simulation in subsection 5.4.2 is shown in Figure 6-2, in which the signal refers to the quantization error ($d = y_q - y$) and noise refers to the disturbance (v). The signal-to-noise ratio is the ratio between the variance of the quantization error excitation ($d = y_q - y$) and the variance of the disturbance (v).

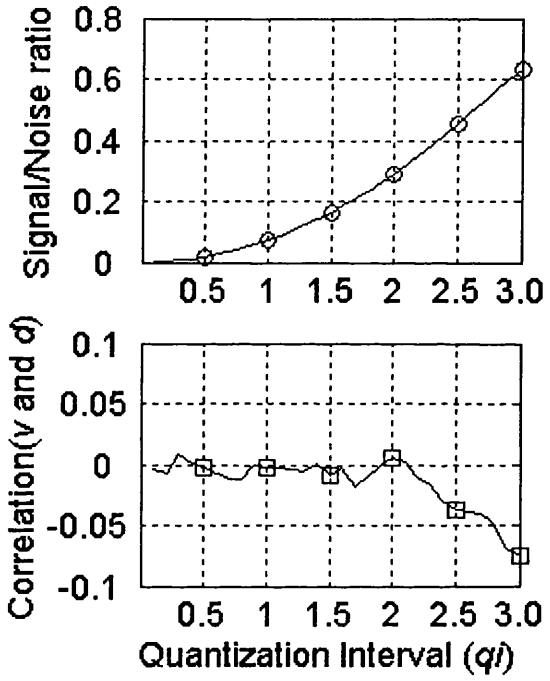


Figure 6-2 *The relationship of the signal-to-noise ratio with qi for the given process and the relationship of the correlation of the quantizer excitation and disturbance with qi.*

The standard deviation of the process output σ_y was 1.373 on a 4 – 20 mA scale when the quantization interval was 0.005 on a 4 – 20 mA scale. From Figure 6-2, the range of the quantization interval from 1.4 to 2.8 is of special interest because the signal-to-noise ratio is large enough for some CLID methods (for example, the two stage method and the two-step method) and the correlation between the excitation and noise/disturbance is negligible. This range (from 1.4 to 2.8) corresponds to σ_y to $2\sigma_y$ approximately. The identification simulation suggests that when the quantization interval is within this range, the identified model accuracy

can be guaranteed. This is the reasoning behind the recommendations given in subsections 4.3.3, 4.3.5 and 5.1.2.

Figure 6-3 is generated from the simulation scheme in Figure 6-1 in which the disturbance is a white noise with variance 1.00 on the 4 – 20 mA scale and the quantization interval is 1.5 times the standard deviation of the process output ($qi = 1.5\sigma_y$). Figure 6-3 is a typical example of unquantized signal, quantized signal and quantization error in the time domain. In the bottom panels of the same figure, the quantization error is expressed in frequency domain. From these two bottom panels, the quantization error is persistently exciting.

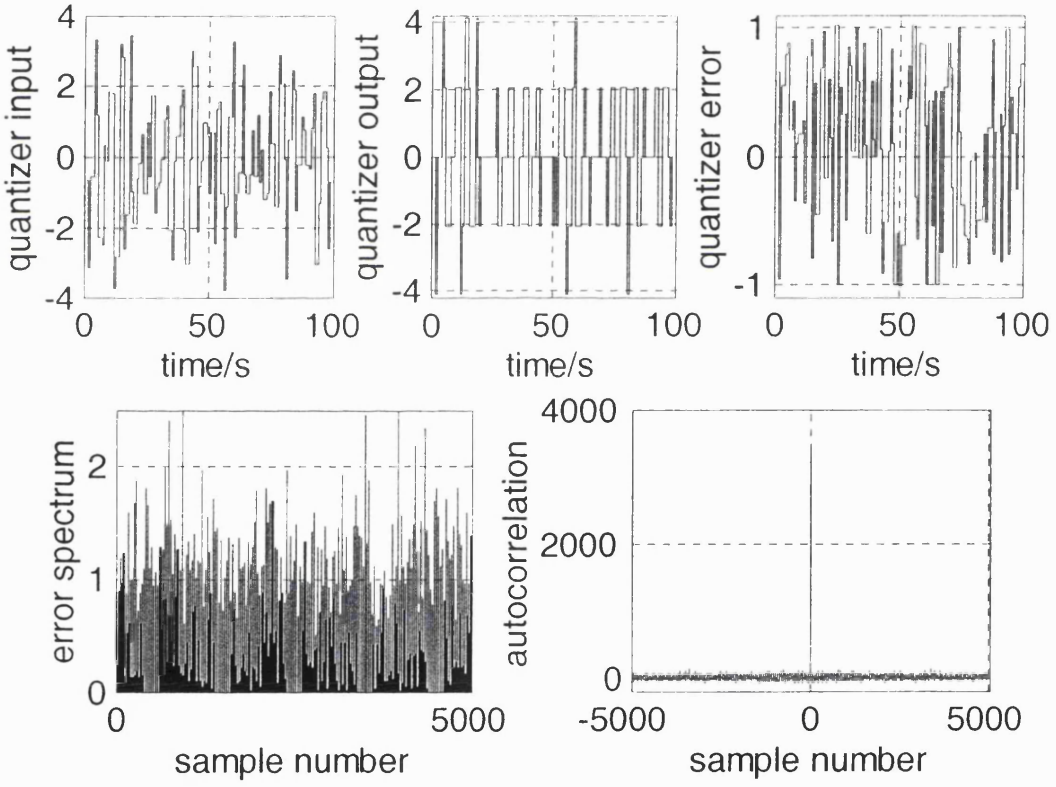


Figure 6-3 *Unquantized signal (left top panel), quantized signal (middle top panel) and quantization error (right top panel), all three panels in the top are in the time domain; the power spectrum of the quantization error (left bottom panel), autocorrelation of the quantization error (right bottom panel), both panels are in the frequency domain.*

6.1.2 Details for further recommendations on quantization interval selection

The results presented in this subsection are from simulations done with Figure 5-6 in subsection 5.4.3, i.e. the discrete ARMAX model with a P-only controller. It is aimed at validating the theory proposed in subsection 4.3.5 of Chapter 4 and explaining the details on further recommendations on quantization interval selection in subsection 5.1.3.

Figure 6-4 indicates how the signal-to-noise ratio and the correlation change with the quantization interval for different controllers. The panels (left to right) in Figure 6-4 correspond to cases 1 to 3 listed in Table 5.1, i.e. for different proportional controller gains. The changes of the CLPA (closed-loop performance assessment) index with prediction horizon are shown in Figure 6-5. According to Thornhill *et al.* (1999), the prediction horizon should be 2 for the three cases. The results are summarised in Table 6.1

	Case 1	Case 2	Case 3
Controller Gain	1.0	0.5	2.0
CLPA index	0.161	0.073	0.594
Minimum Variance	1.584	1.299	2.184
σ_{MV}	1.259	1.140	1.478
σ_y	1.373	1.183	2.317
Range for CLID (*qi*)	1.0 to 2.80	1.30 to 2.40	0.60 to 2.90
Range for CLID (*qi*) expressed in σ_{MV}	$0.80\sigma_{MV}$ to $2.22\sigma_{MV}$	$1.14\sigma_{MV}$ to $2.10\sigma_{MV}$	$0.40\sigma_{MV}$ to $1.96\sigma_{MV}$

Table 6.1 *Results corresponding to simulations of Table 5.1*

In Table 6.1, the standard deviation of the process output σ_y in calculated directly from the measurements. The CLPA index and the minimum variance of the process output are calculated from CLPA theory. σ_{MV} is the square root of the minimum variance of the process output. The row for 'Range for CLID (*qi*)' is found by using different quantization intervals to do identification repeatedly. The row for 'Range for CLID (*qi*) expressed in σ_{MV}' is found by dividing the row for 'Range for CLID (*qi*)' with the corresponding σ_{MV}. Further recommendation will be obtained by finding the common set of the three cases.

From the results in Table 6.1, the further recommendation on selecting the quantization interval when quantization interval is small is from $1.15\sigma_{MV}$ to $1.95\sigma_{MV}$. This is the reasoning behind the recommendations given in subsections 4.3.5, 4.3.6 and 5.1.3.

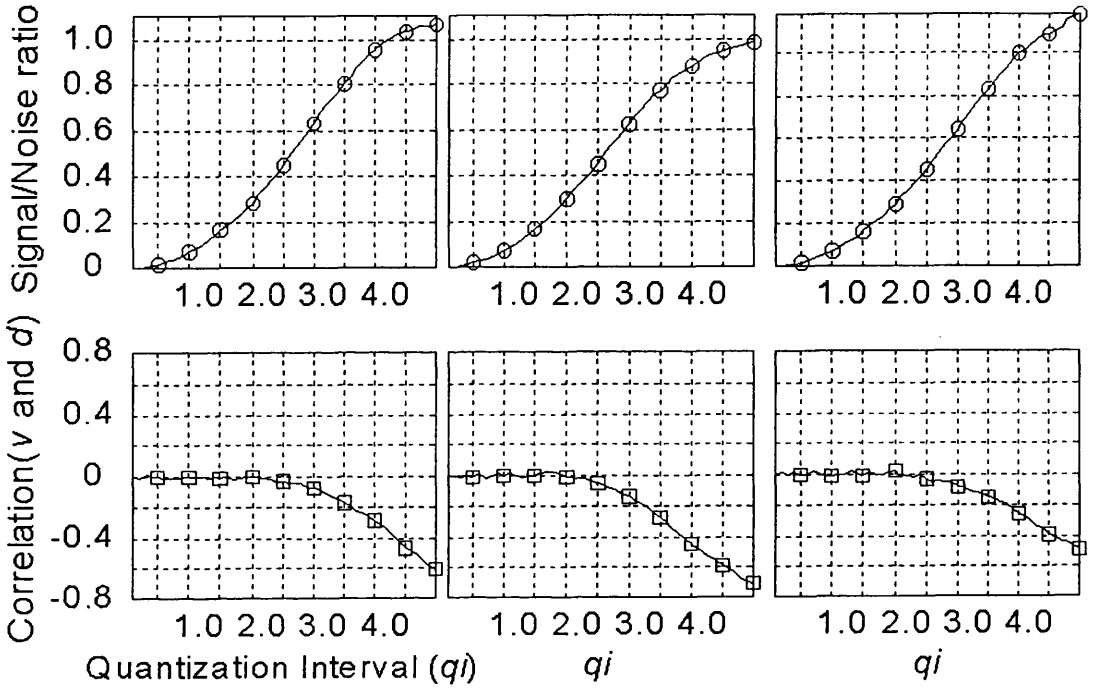


Figure 6-4 *the relationship between the signal-to-noise ratio and qi for the given process and the relationship between the correlation of the quantizer excitation and disturbance and qi for proportional controller with different Gain. Left: Gain 1. Middel:Gain 0.5. Right: Gain 2.0.*

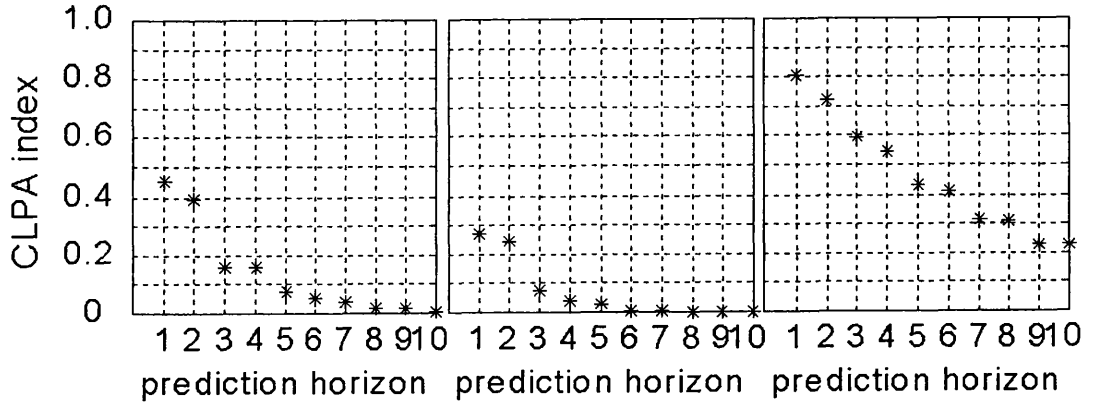


Figure 6-5 *CLPA index for proportional controller with different Gain; Left: Gain 1. Middel:Gain 0.5. Right: Gain 2.0.*

From the CLPA theory, the calculation of σ_{MV} has considered both the process noise and the controller. From the results in Table 6.1, σ_{MV} varies more gently than σ_y for different controllers.

The new recommendation is based on the standard deviation of the minimum variance of the process output σ_{MV}. This is more reasonable since the process noise will contribute to the standard deviation of the process output and the controller will also contribute to the standard deviation of the process output. While the calculation of the standard deviation of the minimum variance of the process output σ_{MV} has included three factors (i.e. the process noise, the controller and the standard deviation σ_y).

6.1.3 Closed-loop identification with external excitation

Results (in the form of Nyquist plots) identified from simulations using CLID with a quantizer (simulation conditions indicated in Table 5.2 in subsection 5.4.4) are shown in Figure 6-6. The panels (top to bottom) in Figure 6-6 correspond to cases 1 to 3 listed in Table 5.2.

From Figure 6-6, the identified model accuracy can be seen qualitatively. With the definition of the worst-case error as the model accuracy measure (subsection 5.3.1), the quantitative results are shown in Table 6.2.

	Huang and Shah (1997)	Van den Hof and Scharama (1993)
Case 1	9.00%	14.44%
Case 2	4.36%	3.93%
Case 3	8.96%	14.50%

Table 6.2 *The worst-case errors for identified models corresponding to simulations in Table 5.2*

From Table 5.2, Case 1 and Case 2 have the same variance of disturbance (0.010). The only difference is the quantization interval applied ($qi = \sigma_y$ in Case 1 and $qi = 1.5\times\sigma_y$ in Case 2). The second panel of Figure 6-6, compared with the first panel of Figure 6-6, shows improved

accuracy of the identified model. This is because the signal-to-noise ratio increased in Case 2 due to the increased quantization interval (*qi*).

Case 1 and Case 3 have different variances of the disturbance a (0.010 *vs* 1.000). However, in both cases, the ratio of the quantization interval and the standard deviation σ_y is the same, that is $qi = \sigma_y$. From the top panel and the bottom panel of Figure 6-6, they have achieved almost the same identification model accuracy. This indicates the proposed scheme is suitable for both small and large disturbances.

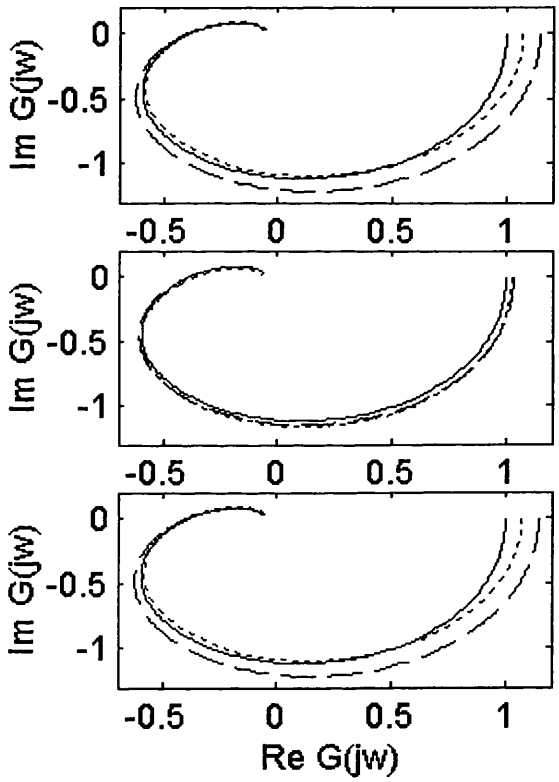


Figure 6-6 *Result for simulations shown in Table 5.2; top panel for Case 1, middle panel for Case 2 and bottom panel for Case3; solid line for true process, dotted line for Huang and Shah (1997), dashed line for Van den Hof and Schrama (1993).*

6.1.4 Conclusion

The simulation examples presented in this section indicated that the proposed identification scheme with a quantizer is successfully implemented for the process expressed in discrete transfer function.

6.2 Continuous transfer function model simulation method

The reasons why this section is necessary are that the whole quantization interval range (i.e. both small quantization interval and large quantization interval) proposed in Figure 4-9 (as well as in Figure 4-15) needs to be emphasized and that the nonlinearity from small to large for the quantization intervals from small to large has to be indicated.

6.2.1 Closed-loop identification with external excitation using a quantizer

The results in this subsection correspond to the simulation designed in subsection 5.5.2.

The standard deviation of the process output is σ_y=0.32 when the quantization interval is 0.005. According to the CLPA theory, the standard deviation of the process output in minimum variance σ_{MV} = 0.32. The quantization interval used is 0.50. The two-stage method (Van den Hof and Schrama, 1993) was used for identification. The identification result is shown in Figure 6-7. The worst-case error of the identified result from zero frequency to the critical frequency is 46.27%.

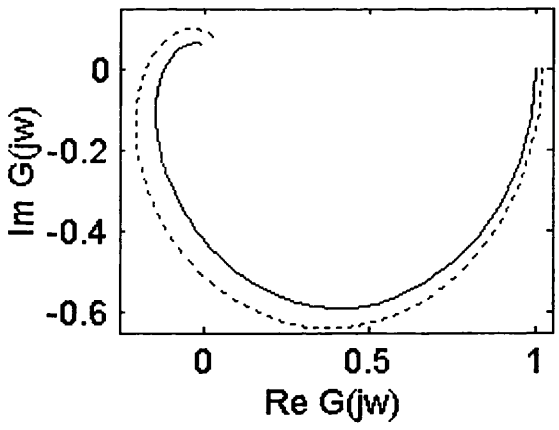


Figure 6-7 Result for closed loop identification with a quantizer for the continuous transfer function: qi is 1.5 times the standarddeviaition of the process output

6.2.2 Relay-based identification with a quantizer

The results in this subsection correspond to the simulation designed in subsection 5.5.3. The standard deviation of the process output is σ_y = 0.32 when the quantization interval is 0.005.

The standard deviation of the process output in minimum variance σ_{MV} = 0.32 by the calculation of CLPA theory. According to the suggestions in Chapter 4, a quantization interval equal to $4\sigma_y$ was used. The identification result is shown in Figure 6-8.

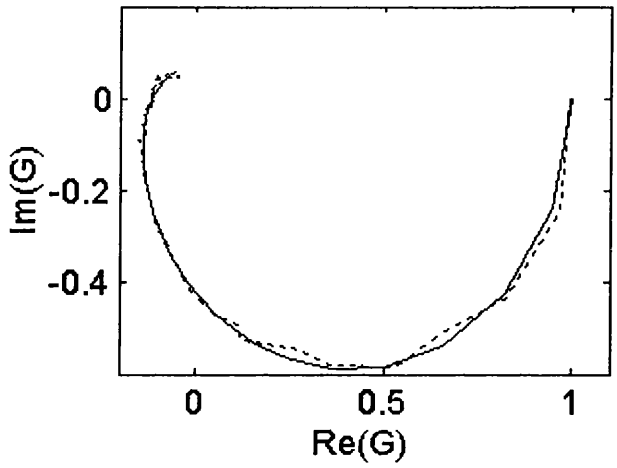


Figure 6-8 Result of relay-based identification with a quantizer for the continuous transfer function process.

The worst-case error of identified result is 35.5%, which occurs near the crossover frequency. The accurate result from the relay-based identification indicates that the suggestion of $4\sigma_y$ for the quantization interval is reasonable.

6.2.3 Nonlinearity test

The results in this subsection correspond to the simulation designed in subsection 5.5.4. This subsection presents the nonlinearity index changes for the two simulation examples presented in subsection 5.5.2 and 5.5.3 in which the quantization intervals are 0.48 and 1.28 respectively (i.e. $1.5\sigma_y$ and $4\sigma_y$ respectively). Table 6.3 gives the nonlinearity test result for the data sets discussed above.

	unquantized	Quantized
Quantization interval 0.48	0.81	3.37
Quantization interval 1.28	0.73	5.02

Table 6.3 *Nonlinearity index for unquantized and quantized signals*

6.2.4 Conclusion

The simulation examples presented in this section indicate that the proposed identification scheme with a quantizer is successfully implemented in the process of the continuous transfer function model.

6.3 Pilot plant model simulation

The arrangement of the pilot plant simulation examples of this section is to validate the whole quantization interval range (small and large quantization interval) proposed in Figure 4-9 (as well as in Figure 4-15) with a real life simulation model instead of the transfer functions used before.

6.3.1 Open loop test

The results in this subsection correspond to the simulations designed in subsection 5.6.2. The open loop identification results of the temperature loop are shown in Figure 6-9. The left panel is the result when the test signal is RBS. It is expressed as a Nyquist plot. The right panel is the unit amplitude step test result.

There is a slight difference in the gain from the two figures. This is due to step test signal is one way (i.e. positive step), while the RBS signal is random with values 1 or –1 and the temperature response has a slightly nonlinear (asymmetric) response.

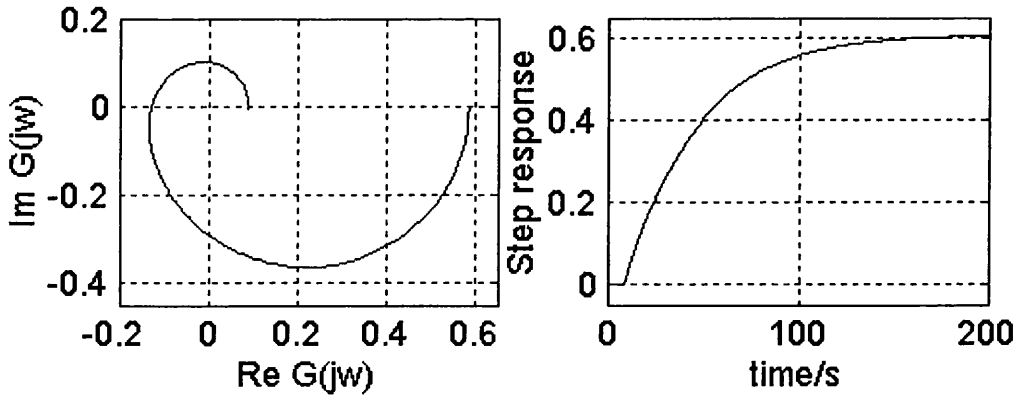


Figure 6-9 The open loop identification results of the temperature loop by simulation model; left: RBS test result; right: step test result.

6.3.2 Closed-loop identification with external excitation using a quantizer

The results in this subsection correspond to the simulations designed in subsection 5.6.3. The standard deviation of the process output is σ_y=0.552 on the 4 – 20 mA scale when the quantization interval is 0.005. According to the CLPA theory, the standard deviation of minimum variance for the process output is σ_{MV} = 0.519. Three cases for different quantization intervals were designed. The two-stage method was used for identification. The open loop identification result with RBS in subsection 6.3.1 is viewed as the true process. The results are shown in Figure 6-10 and Table 6.4

	Case 1	Case 2	Case 3
Quantization interval (*qi*)	0.61	0.81	1.05
Equivalent to σ_y	1.10	1.47	1.9
Worst case error	29.46%	12.49%	18.17%

Table 6.4 *Conditions corresponding to simulation results in Figure 6-10*

From Figure 6-10 and Table 6.4, the errors show that the results of Case 1 and Case 3 are inferior to that of Case 2. In Case 1, the signal-to-noise ratio is lower than that in Case2 because the quantization interval used in Case 1 is much smaller than in Case 2. In Case 3, the correlation is more than that in Case 2 because of the larger quantization interval. Therefore, there is a tradeoff in choosing the quantization interval considering the signal-to-noise ratio and the correlation.

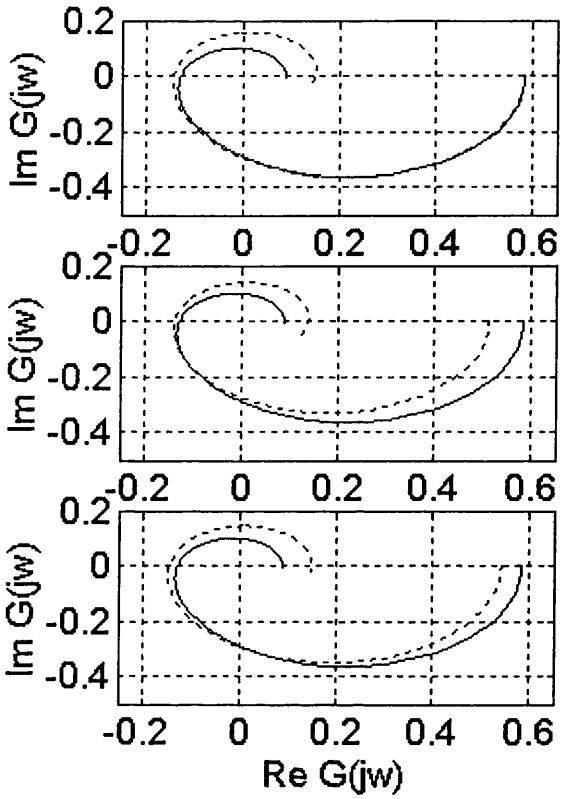


Figure 6-10 *Results for closed loop identification with a quantizer for the temperature loop by simulation model; top: qi is 1.10 times the standard deviation; middle: qi is 1.47 times the standard deviation; bottom: qi is 1.9 times the standard deviation;*

6.3.3 Relay-based identification with a quantizer

The results in this subsection correspond to the simulation designed in subsection 5.6.4. The standard deviation of the process output is σ_y = 0.5522 when the quantization interval is 0.005. The standard deviation of minimum variance for the process output σ_{MV} = 0.5187 by the calculation of CLPA theory. According to the suggestions in Chapter 4, a quantization interval equal to $4\sigma_y$ was used. The identification result is shown in Figure 6-11 in which the open loop identification result with RBS in subsection 6.3.1 is viewed as the true process.

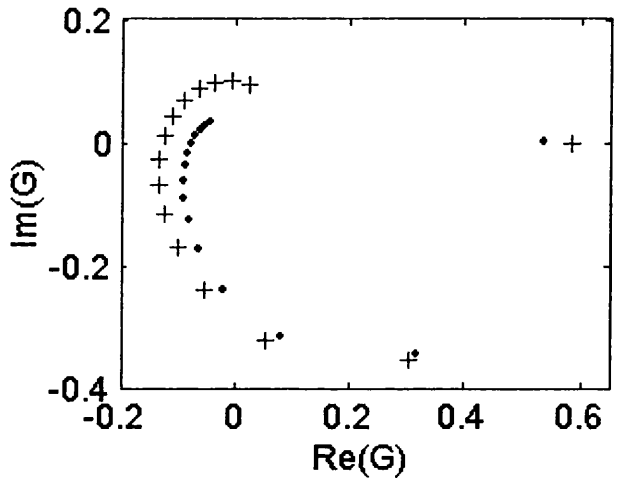


Figure 6-11 Result of relay-based identification with a quantizer for the temperature loop in the tank simulation model (point representss true process and plus represents estimates).

The worst-case error of the identified result compared with the open-loop RBS result is 64% around the critical frequency. It is much worse than the continuous transfer function model simulation.

The differences between the tank simulation model simulation and the transfer function model simulation are due to the following two points. In the tank simulation model simulation, measures have to be taken to prevent load disturbance. In this simulation, the first period of the simulation is the time it takes the system to reach steady-state. The second point is that moderate non-linearity exists in the tank simulation model. This generally influences the identification results. Using a relatively small quantization interval is helpful in avoiding the influence of non-linearity.

6.3.4 Conclusion

The simulation examples presented in this section indicate that the proposed identification scheme with a quantizer is successfully implemented in the temperature loop of the tank simulation model.

6.4 Pilot plant experiment demonstration

The purpose of this section is to demonstrate CLID with a quantizer with a real life pilot plant.

6.4.1 Open loop test

The results in this subsection correspond to the simulations designed in subsection 5.7.1. The left panel of Figure 6-12 is the open loop identification result when the test signal is RBS with the whole frequency band and the amplitude from –0.6 to 0.6. It is expressed as a Nyquist plot. The right panel of Figure 6-12 is the open loop unit amplitude step test result. It is the average of two identical step tests. The step test result is noisy. This is why the step response of the RBS test result is shown again in the right panel of Figure 6-12. From Figure 6-12, it can be seen that the two open loop test results are consistent.

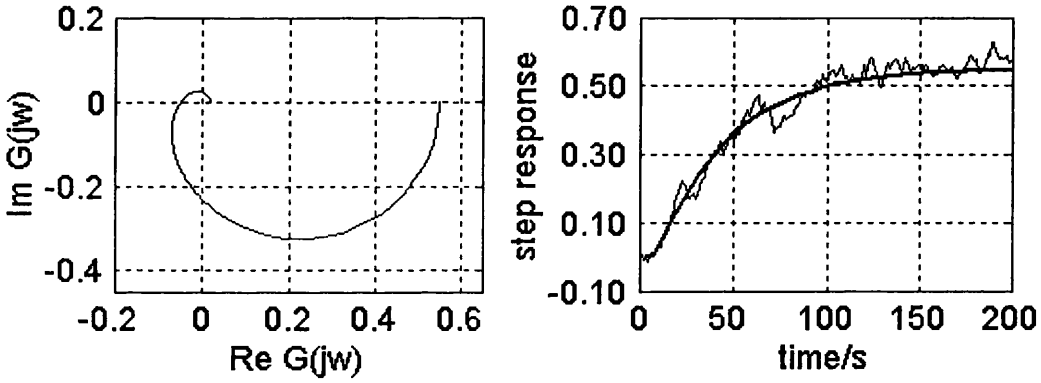


Figure 6-12 The open loop identification results of the temperature loop by experiment; left: RBS test result; right: step test result (thin) and step response identified from the RBS test result (thick).

6.4.2 Closed-loop identification with external excitation using a quantizer

The results in this subsection correspond to the simulations designed in subsection 5.7.2 (i.e. the temperature process of the tank with a PI controller).

Case 1 hybrid experiment

Figure 6-14 shows the result of the closed-loop identification with a quantizer method in which the real tank was used and the disturbance was generated by the computer in the experiment. The variance of the disturbance (a in Figure 5-12) applied in the experiment is 0.25. The

quantization interval used is two times the standard deviation of the process output σ_y = 0.55. That is, $qi = 2\sigma_y = 1.10$. The two-stage method was used for identification.

From the result shown in Figure 6-13, the closed-loop identification result is matched with the open loop result in both low and the medium frequencies. In the high frequency range, the match is not so good. However, commonly in process control, the most important range for identification is the frequency band from zero to crossover frequency. The worst case error of the identified result is 28.98%.

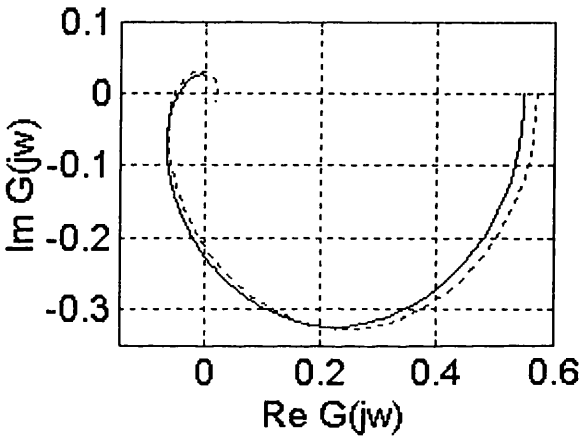


Figure 6-13 *Results for closed loop identification with a quantizer for the temperature loop by hybrid experiment; solid:the open loop RBS test result in subsection 6.4.1; dotted: qi is 2 times the standard deviation for the CLID with a quantizer.*

Case 2 natural noise experiment

Figure 6-14 is the result of the closed-loop identification with a quantizer method in which the real tank was used and the disturbance was natural bubbles in the experiment. The standard deviation of the process output σ_y is calculated to be 0.0657. The quantization interval used is 1.6 times the standard deviation of the process output σ_y. That is $qi = 1.6\sigma_y = 0.105$. The two-stage method was used for identification.

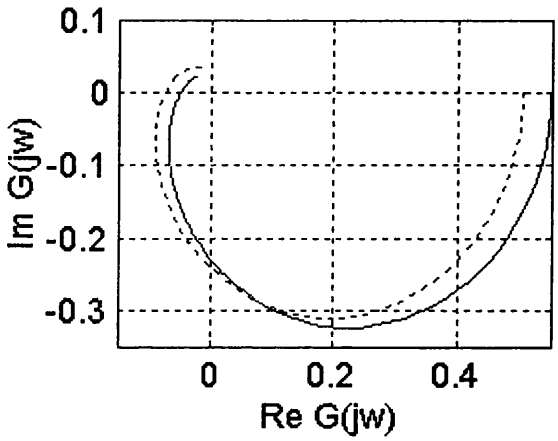


Figure 6-14 Results for closed loop identification with a quantizer for the temperature loop by natural bubble experiment; solid:the open loop RBS test result in subsection 6.4.1; dotted: qi is 1.6 times the standard deviation for the CLID with a quantizer.

Compared with the accuracy in Figure 6-13, the closed-loop identification result with the natural bubbles as shown in Figure 6-14 is slightly worse. The worst case error of the identified result is 54.26%. The reason why the result of Case 2 is worse than the result of Case 1 might be that the natural bubble disturbance applied in Case 2 is not absolutely white noise, while the disturbance applied in Case 2 is white noise generated by the computer which is an ideal case.

6.4.3 Relay-based identification

The relay-based identification method by Wang *et al.* (1997b) needs a sampling rate of at least 0.1 second. The fastest sampling rate in the pilot plant in the University of Alberta is 1 second. Therefore, the method cannot be implemented in experimentation.

6.4.4 Conclusion

The real life experiments presented in this section indicate the validity of the method of closed-loop identification with a quantizer. Because the sampling rate of the pilot plant is not sufficiently fast, the relay-based identification method by Wang *et al.* (1997b) cannot be demonstrated.

6.5 Summary of simulations and experimentations

The simulation examples presented in sections 6.1, 6.2 and 6.3 and the experimentation presented in section 6.4 are in agreement with the theory proposed in Chapter 4. They are summarized in the following Table 6-5 for convenience of reader.

Experimental Condition			Aim	Section	Conclusion
ARMAX Model (equation 5.17)	Unit variance white noise	Qantization interval (*qi*) from 0.1 to 3.0	To show quantization error signal characteristicsq	5.4.2 and 6.1.1	With the increase of qi, S/N increases and correlation becomes more obvious. Useful *qi* range for ID: σ_y to $2\sigma_y$
ARMAX Model (equation 5.17)	Unit variance white noise	Different controller gain used(1.0, 0.5 and 2.0 respectively)	Further recommendation on quantization interval selection	5.4.3 and 6.1.2	σ_{MV} is a better measure for specifying the quantization interval for CLID than σ_y
ARMAX Model (equation 5.17)	White noise with variance 0.01 and 1.0	Different *qi* used (σ_y and $1.5\sigma_y$)	To evaluate the proposed theory for small *qi*	5.4.4 and 6.1.3	useful *qi* range for CLID external excitation: σ_y to $2\sigma_y$
Continuous transfer function model (5.18)	White noise with variance 0.10	$1.56\sigma_y$ quantization interval used	To evaluate the proposed theory for small *qi*	5.5.2 and 6.2.1	The proposed theory on CLID with a quantizer for small *qi* is correct

Continuous transfer function model (5.18)	White noise with variance 0.10	$4\sigma_y$ quantization interval used	To evaluate the proposed theory for large *qi*	5.5.3 and 6.2.2	The proposed theory on CLID with a quantizer for large *qi* is correct
Continuous transfer function model (5.18)	White noise with variance 0.10	Different *qi* used ($1.5\sigma_y$ and $4\sigma_y$)	To show nonlinearity change for small and large *qi*	5.5.4 and 6.2.3	Nonlinearity index increase with *qi*
Pilot plant experimentation	Computer Generated White noise vs Natural bubble	Different *qi* used ($2\sigma_y$ and $1.60\sigma_y$ respectively)	To evaluate the proposed theory for small *qi*	5.7.2 and 6.4.2	Natural bubble disturbance is not so ideal as Computer generated white noise

Table 6-5 *Summary of simulations and experimentations*

6.6 Choice of the CLID methods based on quantization

From theory proposed in Chapter 4 and simulations and experimentation presented in this chapter, it is correct to say that for a given situation when σ_y has been obtained, the identification experiment can be formulated with any quantization interval (*qi*) from σ_y to $2\sigma_y$ for closed loop identification with external excitation and any quantization interval (*qi*) above $4\sigma_y$ for relay-based identification.

As discussed in section 1.3, the motivation of this thesis is to find a novel way to do CLID with as little disruption to the process as possible. By comparison, closed loop identification with external excitation should always be chosen for less disruption since it always choose smaller quantization interval for a given process. Practically, this is not the case.

The insight from the simulation examples and experimentation is that: When the noise is small, it is preferred to use relay-based identification; when the noise is large, it is preferred to use

closed loop identification with external excitation. In practice, one cannot judge the noise size itself. The noise will be reflected in the process output. Therefore, the criterion could be that the case in which the variance of the process output is less than 0.1 on a 4 – 20 mA scale will be viewed as small noise and that the case in which the variance of the process output is equal to and greater than 0.1 on a 4 – 20 mA scale will be viewed as large noise.

The reasoning for adopting this criterion is analysed as follows. When the noise is small, the burden of filtering the noise is not heavy and the quantization interval required is still reasonable compared with the relay identification. At this moment, the outliers in the measurements might make the quantization error not white in the industrial experiment. This point is obviously illustrated from the experimental evaluation presented in the two cases in subsection 6.4.2.

6.7 Conclusions

Based on the simulation and experiment results presented in the above sections of this chapter. The conclusions are:

- The quantization interval range in Figure 4-9 (as well as in Figure 4-15) indicates that closed-loop identification with external excitation method can be used when the quantization interval is small and relay-based identification can be used when the quantization interval is large. This conclusion has been demonstrated by using simulations and experiments.
- The insights from this chapter are: In order to minimize the disruption to the system, when the noise is small, relay-based identification is recommended; when the noise is large, the closed-loop identification with external excitation method is recommended. The criterion to judge whether the noise is small or large is the variance of the process output in which small noise corresponds to σ_y^2 less than 0.1 and large noise corresponds to σ_y^2 greater than 0.1.

CHAPTER 7

TIME SERIES RECONSTRUCTION FROM QUANTIZED MEASUREMENTS

This is a self-contained chapter. This chapter describes the implementation of a Quantization Regression (QR) algorithm that generates nonlinear estimates from quantized measurements. The algorithm, based on Ziskand and Hertz (1993), is a combination of the 'Gaussian Fit' scheme of Curry (1970) and the expectation-maximization (EM) algorithm of Dempster *et al* (1977). The efforts are then directed to showing how to use the algorithm. The QR algorithm is then compared with the Linear Least Square (LLS) method and the Kalman smoothing method (the same formulation as the QR algorithm but without the Gaussian Fit scheme) in the performance of reconstruction from the coarsely quantized signal. The purpose of the comparisons is to find out which element plays an important role in signal recovery. From the simulation examples and experimentation presented, the conclusion is that the QR algorithm is a more accurate method. This allows the retrieval of the underlying information from quantized signals.

As explained later in this chapter, the QR algorithm was implemented successfully but was not in the end used in the closed-loop identification procedure because its large computation time makes it difficult to use on-line.

7.1 Introduction

This section begins by presenting the motivation. Then the progress of recovering from the quantization measurement is reviewed. Finally, the novel contribution of the author in this chapter is stated.

7.1.1 Motivation

Signal processing has played a major role in diverse fields such as speech and data communication, oil exploration, instrumentation, acoustics, sonar and radar. Time series analysis, a branch of signal processing, is the procedure of fitting stochastic models to the observations available at discrete points in time. There are many time series models, such as

autoregressive (AR), moving average (MA) and autoregressive and moving average (ARMA). Among all these models, AR models are much easier to fit than MA or ARMA models. However, AR model may require extra parameters to get a sufficiently good representation of the data (Chatfield, 1989). Suppose that $\{e(n)\}_{n=1}^{N}$ is a purely random process with mean zero and variance σ_e^2. Then, a process $\{x(n)\}_{n=1}^{N}$ is said to be an autoregressive process of order m if

$$x(n) = a_1 \cdot s(n-1) + a_2 \cdot s(n-2) + \cdots + a_m \cdot s(n-m) + e(n) \quad (7.1)$$

and $a_m \neq 0$. A great deal of data in engineering occurs in the form of time series where observations are dependent and where the nature of this dependence is of interest in itself. For example, Desborough and Harris (1992) proposed a linear regression approach to reconstruct from routine closed loop process data where the purpose is to estimate the normalized performance index. In this case, an AR model was implemented.

Thornhill *et al.* (1999) observed that quantization of the measurements influenced the normalized performance index, and thus motivated the need for an algorithm to accurately recover the underlying signal from quantized measurements.

7.1.2 Recent progress

Ziskand and Hertz (1993) proposed a novel algorithm to estimate the coefficients of multiple AR processes from a coarsely quantized signal. Their implementation and simulation examples illustrated the case of one-bit (two-level) quantization. The algorithm was based on the following two developments.

Dempster *et al.* (1977) presented a broadly applicable algorithm for computing maximum likelihood estimates from "incomplete" data, i.e. data having a many-to-one mapping in the measurement function. Quantized process data are incomplete because many values of the underlying signal map to each quantizer level. The algorithm by Dempster *et al.* (1977) includes two steps: the first expectation, the second maximization, and is called the EM algorithm. Shumway and Stoffer (1982) proposed an approach to smoothing and forecasting for time series with missing observations. The EM algorithm was used in conjunction with the

conventional Kalman smoothed estimators to derive a recursive procedure for estimating the parameters by maximum likelihood.

The 'Gaussian Fit' algorithm was proposed in Curry (1970) for a discrete-time nonlinear filter that recursively fits a Gaussian distribution to the first two moments of the conditional distribution of a system state vector. The primary advantages of the Gaussian fit algorithm are that it is relatively easy to compute, that it can handle nonstationary data as easily as stationary data, and that its general operation is independent of the quantization scheme used. The disadvantages are that it requires more computation than the optimal linear filter and that it can be applied only to Gauss-Markov processes. It is applicable to quantized process data because quantized process data can be expressed as a Gauss-Markov process in state-space form (as shown in the section 7.2).

In the last two decades, the EM algorithm has become an important general technique for finding maximum likelihood estimates (MLE) from incomplete or missing data, especially when the missing mechanism is unknown (Little and Rubin, 1986). In this chapter, quantized process data can be viewed as incomplete or missing data with unknown missing mechanism because many values of the underlying signal map to each quantizer level.

Meng and Rubin (1991) proposed a Supplemented EM (SEM) algorithm, which provides the asymptotic variance-covariance for parameters (e.g., standard errors) based on the second derivatives of the observed-data log-likelihood. Quasi-Newton methods were used in Jamshidian and Jennrich (1997) to speed up the convergence rate of the EM algorithm. In the context of incomplete or missing data, Akaike Information criterion (AIC) is often used to determine model order. Cavanaugh and Shumway (1998) proposed a criterion called AICcd, where the notation 'cd' stands for 'complete data', which is used to determine the model order more accurately. These developments will be useful in future version of Quantization Regression (QR) algorithm.

7.1.3 Novel contribution

The algorithm presented by Ziskand and Hertz (1993) estimated the coefficients of several superimposed multiple AR signals of known model order using a two-level quantizer. The several signals represented sinusoids at different frequencies. The contribution of this chapter is

the extension of quantized regression for one AR signal embedded in noise to multiple quantizer levels. Other improvements will be listed for clarity.

- Some typing errors and concept errors in the Ziskand and Hertz (1993) paper have been corrected. These will be presented in subsection 7.2.3.
- The stopping criterion of the iteration loop is a trade-off between accuracy and computation. In Ziskand and Hertz (1993), the number of iterations was specified according to different occasions judged by the researchers. Here, the criterion that the innovations form of Log Likelihood function be stable is used.
- The model order was assumed to be known in Ziskand and Hertz (1993). Here, the model order m is determined objectively by the Akaike Information Criterion (AIC). This will be presented in subsection 7.3.1.
- The method has been used to determine the practical default settings for parameters such as the ensemble length of the test data necessary to this algorithm.
- Simulation examples and experimental data have been used to demonstrate the superiority of this algorithm compared to other methods.

7.2 Quantized regression (QR) algorithm

Quantized autoregressive time series are the subject of this work. The problem statement of the algorithm is given in this section. Key elements of the algorithm are then described. Improvements are described in detail at the end.

7.2.1 Problem description

The time series (7.1) is rewritten in state-space form as follows,

$$S(n) = \phi \cdot S(n-1) + w(n) \tag{7.2}$$

$$x(n) = h \cdot S(n) + v(n) \tag{7.3}$$

$$z(n) = g(x(n)) \tag{7.4}$$

where $S(n) = (s(n-m+1),\cdots,s(n))^T$ is a state vector, and

$$\phi = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \cdot\cdot & \cdot\cdot & \\ & & & 0 & 1 \\ a_m & \cdot\cdot & \cdot\cdot & a_2 & a_1 \end{pmatrix}$$

is the transition matrix. In this matrix, elements in the last row $a_1, a_2, \ldots, a_m$ are the autoregressive coefficients,

h is the observation matrix $h = [h_l]$

$$h_l = \begin{cases} h_l = 1, & l = m \\ h_l = 0, & otherwise \end{cases}$$

$w(n)$ and $v(n)$ are all white measurement noise, with mean zero and variance σ_w^2, σ_0^2 respectively. $v(n)$ is generally termed the *observation noise.* More importantly, they are assumed to be independent from each other.

$\{z(n)\}_{n=1}^{N}$ are the quantized measurements of the process, and

$\{x(n)\}_{n=1}^{N}$ are the N samples from an autoregressive process

g is the non-linear quantizer function whose input is the AR signal plus noise.

In standard textbooks (for example, Shumway and Stoffer, 2000), (7.2) is named the *state equation* and (7.3) is called the *observation equation.*

Problem statement

The quantization interval is the distance between quantizer levels. A uniform quantizer refers to a quantizer that has the same quantization interval. The quantizer used in this chapter will be a uniform quantizer (even though the algorithm is suitable for both uniform and non-uniform quantizers, from 1 bit to 12 bit). The problem is to determine the model order, m, to estimate the AR coefficients $a_1, a_2, \ldots, a_m$ and to recover the underlying signal $\{s(n)\}_{n=1}^{N}$.

7.2.2 Key elements in the QR algorithm

Kalman filter

When the system is rewritten in state-space as in (7.1) and (7.2), the practical aim is to estimate the underlying unobserved s_t, given the data $x_n = \{x_1, x_2, \ldots, x_n\}$.

- If $n < t$, the problem is called *forcasting* or *prediction.*
- If $n = t$, the problem is called *filterting.*
- If $n > t$, the problem is called *smoothing.*

Kalman filter gives the filtering and forecasting equations. The advantage of the Kalman filter is that it specifies the updating from s_{t-1}^{t-1} to s_t^t once a new observation $x(n)$ is obtained. It does not need to reprocess the entire data set $\{x_1, x_2, ..., x_n\}$. For the sytem described by (7.1) and (7.2), when the initial conditions have been given, such as mean $\mu = s_0^0$ and covariance $\sum_0 = P_0^0$, the Kalman filter for $t = 1,2,...,n$ will be given without proof. The textbook by Shumway and Stoffer (2000) gives a detailed description.

$$s_t^{t-1} = \phi s_{t-1}^{t-1} \tag{7.5}$$

$$P_t^{t-1} = \phi P_{t-1}^{t-1} \phi^T + \sigma_w^2 \tag{7.6}$$

$$K_t = P_t^{t-1} h^T [h P_{t-1}^t h^T + \sigma_0^2]^{-1} \tag{7.7}$$

$$s_t^t = s_t^{t-1} + K_t (x_t - h s_t^{t-1}) \tag{7.7}$$

$$P_t^t = [\mathrm{I} - K_t h] P_t^{t-1} \tag{7.8}$$

Where K_t is called the Kalman gain, P_t^t is the conditional error covariance.

Gaussian Fit

The 'Gaussian Fit' algorithm was proposed in Curry (1970) for a discrete-time nonlinear filter that recursively fits a Gaussian distribution to the first two moments of the conditional distribution of a state vector. The scheme is based on the assumption that the conditional probability density function of the state $s(n)$ given the quantized measurements $z(n)$ is Gaussian, i.e.

$$f(s(n) \mid z_{n-1}) \sim N(s_n^{n-1}, P_n^{n-1}) \tag{7.9}$$

The scalar Guassian case is given in the following according to Curry (1970). Each quantized scalar measurement requires two numbers as the filter input. These numbers are the bounds of the quantization interval in which the measurement lies. Here $a \le z \le b$ $(a<b)$ is assumed.

The scaling factor can be calculated by

$$P(z \in (a,b)) = \frac{1}{2}[erf(\frac{b-\bar{x}}{2^{1/2}\sigma_x}) - erf(\frac{a-\bar{x}}{2^{1/2}\sigma_x})] \tag{7.10}$$

where *erf(X)* is the error function for each element of X. X must be real. The error function is defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt \tag{7.11}$$

The conditional mean can be calculated by

$$E(x(n) \mid z_n) = \bar{x} + \frac{\sigma_x}{P(z \in (a,b))}[\frac{\exp\{-\frac{1}{2}((a-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}} - \frac{\exp\{-\frac{1}{2}(b-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}}] \tag{7.12}$$

The conditional covariance can be calculated by

$$\begin{aligned}\mathrm{cov}(x(n) \mid z_n) = \sigma_x^2[1 + (\frac{a-\bar{x}}{\sigma_x})\frac{\exp\{-\frac{1}{2}((a-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}P(z \in (a,b))} - (\frac{b-\bar{x}}{\sigma_x})\frac{\exp\{-\frac{1}{2}((b-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}P(z \in (a,b))} \\ - \frac{1}{P(z \in (a,b))^2}(\frac{\exp\{-\frac{1}{2}((a-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}} - \frac{\exp\{-\frac{1}{2}((b-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}})^2]\end{aligned} \tag{7.13}$$

The QR algorithm

The QR algorithm adopted the EM framework of Dempster *et al.* (1977). The framework was further developed in Shumway and Stoffer (1982). The difference between this framework and the conventional 'innovations' likelihood is that the latter only requires the recursions, while the former requires both forward recursions and backward recursions. Essentially the framework is multivariate regression calculations (Shumway and Stoffer, 1982). In the new

algorithm, the end result of interest is the smoothed series and its covariance matrix $S_n^N, P_n^N, P_{n,n-1}^N$.

An important note is that the problem stated in subsection 7.2.1 is slightly different from the generic problem. What can be measured here are the quantized measurements $\{z(n)\}_{n=1}^N$ instead of $\{x(n)\}_{n=1}^N$ as usual. This determines the peculiarity that the expectation $E(x(n) \mid z_n)$ is first calculated through the 'Gaussian Fit' scheme, and then used to substitute x_t as in (7.7). Similarly, $\mathrm{cov}(x(n) \mid z_n)$ is first calculated through the 'Gaussian Fit' scheme, and then used as an additional item added in (7.8).

The Quantization Regression (QR) algorithm is iterative and includes two main conceptual steps (Ziskand and Hertz, 1993):

Step 1: A modified Kalman smoothing algorithm finds smoothed estimates and their covariance matrix $s_n^N, P_n^N, P_{n,n-1}^N$. At step n the key calculations are:

$$s_n^n = s_n^{n-1} + K_n((E(x(n)) \mid z_n) - hs_n^{n-1}) \tag{7.14}$$

$$P_n^n = P_n^{n-1} - K_n h P_n^{n-1} + k_n \mathrm{cov}\{x(n) \mid z_n) K_n{}^T \tag{7.15}$$

where s_n^{n-1} and P_n^{n-1} are one step ahead predictions of the state vector and its variance, z_n the sequence of quantized measurements and K_n is the Kalman gain. A conventional Kalman Filter would use the quantized measurement z_n directly, but the modified algorithm calculates the expected value of $x(n)$ given that the quantized measurement falls within the observed quantization interval. The expectation is computed using the Gaussian Fit approximation (Curry, 1970). The last term in the calculation for P_n^n captures the inflation in the variance of the estimate caused by quantization.

Step 2: The likelihood function is maximised by iterative adjustment of the AR coefficients $a_1, a_2, ..., a_m$, the estimated signal and σ_w^2 and σ_o^2 (with information obtained in step (1)).

7.2.3 Improvement of the algorithm

Typing errors found in Ziskand and Hertz (1993) were corrected as follows where the underline indicates the altered terms,

$$K_n = P_n^{n-1} h^T \underline{(\sigma_x^2)}^{-1} \tag{7.16}$$

$$P_{n-1}^N = \underline{P_{n-1}^{n-1}} + J_{n-1}(P_n^N - P_n^{n-1}) J_{n-1}^T \tag{7.17}$$

$$P_{n-1,n-2}^N = P_{n-1}^{n-1} J_{n-2}^T + \underline{J_{n-1}}(P_{n,n-1}^N - \phi(r) P_{n-1}^{n-1}) \underline{J_{n-2}^T} \tag{7.18}$$

and the loop in the lag-one covariance calculation was for $n = N, N-1, \ldots, \underline{2}$.

The following improvements were made:

(1) σ_o^2 was updated during the iteration rather than being taken as known.

(2) The innovative form of the log liklihood function (Shumway and Stoffer, 1982) was used as a stopping criterion.

$$\Delta(\log L) = -\frac{1}{2}\sum_{n=1}^{N}\log(\sigma_x^2) - \frac{1}{2\sigma_x^2}\sum_{n=1}^{N}(z(n) - hs_n^{n-1})^T (z(n) - hs_n^{n-1}) \tag{7.19}$$

Iterations stopped when the innovation became small.

(3) The initial variance in the r_{th} iteration was fixed and the state vector was initialised thus: $S_o^o(r+1) = S_o^N(r)$.

7.2.4 Matlab implementation

Based on the description above, the modified QR algorithm is listed mathematically in **Appendix A**. The algorithm was implemented in Matlab.

7.3 Use of the QR Algorithm

Determination of the model order is described first. Then the method to determine ensemble length is briefly presented. Emphasis is given to the comparison of another two methods with QR. The performance measure used to measure the quality of fit (degree of recovery of the underlying signal) is given at the end.

7.3.1 Determination of model order

Determining the order m of an autoregressive process is very important. One approach to fit AR process of progressively high order, is to calculate the sum of squared errors (SSE) for each value of m, and to plot this against m (Chatfield 1989). It may be possible to see the value of m where the curve flattens out such that SSE becomes insensitive to increasing m. However, this method does not work here since the aim is to recover the underlying signal, not the quantized observation.

Another approach, especially in the context of maximum likelihood estimation (MLE), is to use Akaike Information criterion (AIC) to determine model order. For the incomplete (missing) data problem, because the EM algorithm utilizes complete-data tools, it may be more convenient to determine model order m based on complete-data quantities rather than analogous incomplete-data quantities.

The complete-data log likelihood can be computed as follows (Ziskand and Hertz, 1993; Shumway and Stoffer, 1982).

$$\begin{aligned}\log L(\theta \mid \theta_i) &= -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}tr\{\Sigma^{-1}[P_0^N + (s_0^N - \mu)(s_0^N - \mu)^T]\} \\ &-\frac{N}{2}\ln\sigma_w^2 - \frac{1}{2\sigma_w^2}[(C_i)_{(m,m)} - (B_i)_{(m,:)}\phi(m,:)^T - \phi_{(m,:)}(B_i)_{(m,:)}{}^T + \phi_{(m,:)}D_i\phi_{(m,:)}{}^T]\} \\ &-\frac{N}{2}\ln\sigma_0^2 - \frac{1}{2\sigma_0^2}\sum_{n=1}^{N}[(z(n) - hs_n^N)(z(n) - hs_n^N)^T + hP_n^N h^T]\end{aligned} \tag{7.20}$$

where

$$B_i = \sum_{n=1}^{N}\{P_{n,n-1}^N + s_n^N(s_{n-1}^N)^T\} \tag{7.21}$$

$$C_i = \sum_{n=1}^{N}\{P_n^N + s_n^N(s_n^N)^T\} \tag{7.22}$$

$$D_i = \sum_{n=1}^{N}\{P_{n-1}^N + s_{n-1}^N(s_{n-1}^N)^T\} \tag{7.23}$$

and $\Sigma = P_0^0$, $|\Sigma|$ means determinant of Σ, $\mu = s_0^0$, $tr\{\}$ means matrix trace.

AIC can be calculated by

$$AIC = -2\log L(\theta \mid \theta_i) + 2m \qquad (7.24)$$

where m is the model order. The model order is the value of m where the largest decrease occurs.

7.3.2 Determination of ensemble length

Theoretically, the variance–covariance matrices for parameters should be known to determine the simulation ensemble length for the QR algorithm. This is not available in the EM algorithm unless SEM is used (Meng and Rubin, 1991). Here different lengths were used in the simulation (details can be referred to in sections 7.4 and 7.5) where the quantizer input is a unit amplitude sinusoid wave with a certain quantization interval, to show the influence of ensemble length.

7.3.3 Methods for comparison — LLS method and Kalman smoothing method

The QR method will be compared with two other methods for recovery performance. One method was a one-step linear least squares estimate (*LLS*) of the AR coefficients. The AR process (7.1) is rewritten in matrix form

$$\begin{pmatrix} s(n) \\ s(n-1) \\ \cdot \\ \cdot \\ s(m+1) \end{pmatrix} = \begin{pmatrix} s(n-1) & \cdot & \cdot & \cdot & s(n-m) \\ s(n-2) & \cdot & \cdot & \cdot & s(n-m-1) \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ s(m) & \cdot & \cdot & \cdot & s(1) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_m \end{pmatrix} + \begin{pmatrix} e(n) \\ e(n-1) \\ \cdot \\ \cdot \\ e(m+1) \end{pmatrix} \qquad (7.25)$$

It can also be written as:

$$Y = X \cdot a + e \qquad (7.26)$$

The parameter estimates can be found by solving the simultaneous linear equations

$$X^T \cdot Y = X^T \cdot X \cdot a \tag{7.27}$$

The AR coefficients are given by

$$a = (X^T \cdot X)^{-1} \cdot X^T \cdot Y \tag{7.28}$$

The above calculation coincides with the Linear Regression approach in Desborough and Harris (1992) using a one-step prediction. It was also used in Thornhill *et al.* (1999) where the adverse influence of quantization was noted.

The other method, termed *Kalman smoothing*, used the same formulation as the QR algorithm but without the Gaussian Fit step. The Kalman smoother equations became:

$$s_n^n = s_n^{n-1} + K_n(z(n) - hs_n^{n-1}) \tag{7.29}$$

$$P_n^n = P_n^{n-1} - K_n h P_n^{n-1} \tag{7.30}$$

7.3.4 Performance Measure

When a model has been fitted to a time series, it is advisable to check whether the model does provide an adequate description of the data. The multiple coefficient of determination R^2 (Scheaffer and McClave, 1995) was used to determine the quality of the fit.

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y(n) - \hat{y}(n))^2}{\sum_{n=1}^{N}(y(n) - \bar{y}(n))^2} = 1 - \frac{SSE}{SS_{yy}} \tag{7.31}$$

where $\hat{y}(n)$ is the reconstruction of a true underlying signal $y(n)$ and $\bar{y}$ is its mean. $R^2 = 0$ implies a complete lack of fit of the model to the data, and $R^2 = 1$ implies a perfect fit, with the model passing through every data point. In other words, the larger the value of R^2, the better the model fits the data.

7.4 Simulation and Experimentation

In this section, simulation examples and experimentation data will be used to demonstrate the goodness of the algorithm.

7.4.1 Simulation Examples

In this subsection, different simulation cases were designed for different purposes:

- Case 1 is designed for the model order determination.
- Case 2 is designed for the determination of ensemble length.
- Case 3 is designed to see whether the algorithm is robust under different initial conditions.
- Case 4 is designed to compare the QR algorithm with other methods in recovery performance.
- Case 5 is designed to see the recovery performance under different quantization intervals.

A unit amplitude sinusoid wave (contaminated with noise or without noise) with quantization interval 0.5 was widely used in the simulation example (repeatedly referred to as the unit amplitude sinusoid wave later). Therefore, the signal used just five quantizer levels and was coarsely quantized. The sine wave had a period of 51.2 samples. The value was selected so that 10 cycles used 512 samples (a power of 2).

Case 1: Determination of Model Order

The unit amplitude sinusoid wave was passed through a quantizer with quantization interval 0.5. The quantized measurements were used to test the algorithm. The simulation data ensemble length used is $N = 1440$. Similarly, the unit amplitude sinusoid wave contaminated with white noise (signal to noise ratio is 1 to 1) was applied for another test of the same type.

Case 2: Determination of Ensemble Length

The unit amplitude sinusoid wave was passed through a quantizer with quantization interval 0.5. The quantized measurements were used to test the algorithm. Different numbers of the quantized measurements were used in the different simulations to show the influence of ensemble length.

Case 3: Robustness of the Algorithm

The unit amplitude sinusoid wave was passed through a quantizer with quantization interval 0.5. The quantized measurements were used to test the algorithm. The simulation data

ensemble length used is fixed $N = 1440$. But an initial condition was changed from the standard condition in each test.

Case 4: Comparison with Other Methods in Recovery

The unit amplitude sinusoid wave was passed through a quantizer with quantization interval 0.5. The quantized measurements were used to test the algorithm. Similarly, the unit amplitude sinusoid wave contaminated with white noise (signal to noise ratio is 1 to 1) was applied for another test of the same type. As described in subsection 7.3.3, QR algorithm was compared with both the Linear Least Square (LLS) method and the Kalman smoothing method (the same formulation as the QR algorithm but without the Gaussian Fit step).

Case 5: Influence of the Quantization Interval

The unit amplitude sinusoid wave was passed through a quantizer with a range of quantization intervals (from 0.1 to 1.0 with increment 0.1). The quantized measurements were used to test the algorithm. Similarly, the unit amplitude sinusoid wave contaminated with white noise (signal to noise ratio is 1 to 1) was applied for another same test. As described in subsection 7.3.3, the QR algorithm was compared with both the Linear Least Square (LLS) method and the Kalman smoothing method.

7.4.2 Experimentation

The measurements were from the transient response of the pH control in a buffered fed-batch yeast fermentation process. The A/D converter of the pH probe quantized them. The aim of the analysis was to recover the underlying smooth transient response of the fermenter pH.

7.5 Results and Discussions

Results from the simulation examples and experimentation data are presented in this section. The discussions are based on the corresponding results.

7.5.1 Simulation examples

Results from the simulation examples and the discussions based on the results are presented in this subsection case by case.

Case 1: Determination of Model Order

The result for the unit amplitude sine wave with quantization interval 0.5 is shown in the left panel of Figure 7-1. Correspondingly, the result for the unit amplitude sine wave contaminated with white noise (signal to noise ratio is 1 to 1) with quantization interval 0.5 is shown in the right panel of Figure 7-1.

The theory described in subsection 7.3.1 is that the model order is the value of m where the largest decrease occurs. Obviously, m = 2 in the former case and m = 5 in the latter case should be an appropriate choice from Figure 7-1.

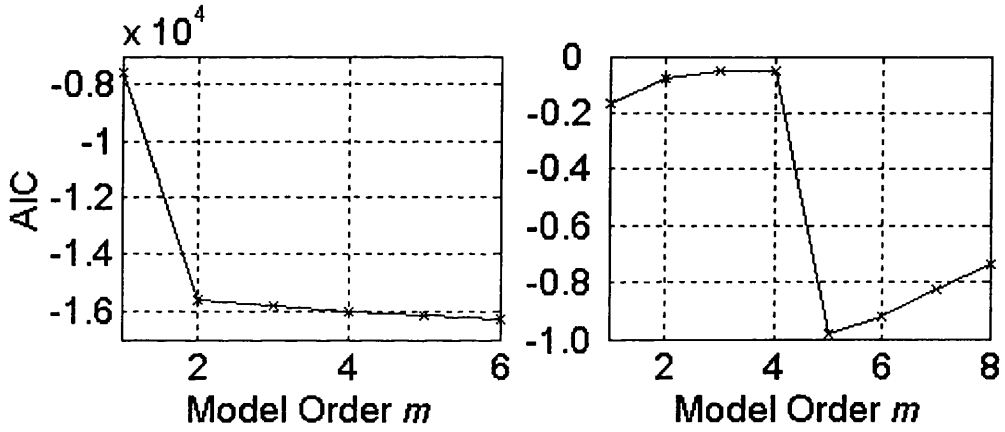


Figure 7-1 *Model order determination for a sine wave with quantization interval 0.5 (left) and for a sine wave contaminated with white noise with quantization interval 0.5 (right)*

Case 2: Determination of Ensemble Length

The unit amplitude sinusoid wave used had a period of 51.2 samples. Using the method of backward differencing the sine wave $\sin(2\pi / 51.2)$ can be expressed as a two term AR series:

$$y(n) = 1.990y(n-1) - 0.998y(n-2)$$

Therefore, the QR algorithm is required to give a model with $m = 2$ and coefficients $a_1 = 1.990$ and $a_2 = -0.998$. The AIC (Case 1 above) indicated that the optimum model

order for the AR time series was indeed $m = 2$. The QR algorithm was used to reconstruct an AR model with two terms and the results were $a_1 = 1.977$ and $a_2 = -0.993$.

When different lengths were used in the simulation, the results changed slightly from test to test. The results are listed in the following Table 7.1.

The choice of ensemble length can be regarded as a trade-off between narrow confidence limits for estimates and loss of information (Desborough and Harris 1992).

Ensemble Length	a_1	a_2	R^2
N=1440	1.977	-0.993	0.9981
N=1280	1.979	-0.994	0.9981
N=1024	1.977	-0.992	0.9980
N= 768	1.974	-0.989	0.9979
N= 512	1.968	-0.984	0.9977
N= 256	1.952	-0.968	0.9976

Table 7.1 *Test Results for the influence of ensemble length*

From Table 7.1, in this example a data ensemble length of 800 gave errors compared to the true values of about 1% in each of the AR coefficients; therefore, it is recommended that the data ensemble length be at least 800 samples.

Case 3: Robustness of the Algorithm

The results as well as the changed initial conditions for different tests are listed in Table 7.2.

Condition Change	Results	Maximum Errors
Standard	AR=-0.9925 1.9772	Standard
Q(0) = 100	AR=-0.9925 1.9772	0.0000
R(0) = 100	AR=-0.9925 1.9772	0.0000
Q(0) = 100; R(0) = 100	AR=-0.9925 1.9772	0.0000
V(0) = $\bar{e}$ (0 . . 0)'	AR=-0.9925 1.9772	0.0000

$P = e^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	AR=-0.9924 1.9771	0.0001
$S = (0 \ . \ . \ 0)'$	AR=-0.9925 1.9772	0.0000

Table 7.2 *Test Result for initial condition influence*

The results in Table 7.2 show that the QR algorithm is robust. It is nearly insensitive to any initial condition change.

Case 4: Comparison with Other Methods in Recovery

The QR algorithm was used to reconstruct an AR model with two terms and the results were compared with the reconstruction of $m = 2$ AR models using Kalman smoothing alone and the one-step linear least squares estimate. The recovery ability is compared in the left panel of Figure 7-2, where it can be seen visually that the QR algorithm provided the best reconstruction of the underlying sine wave. The results with the LLS method were the least satisfactory.

The estimated AR coefficients and R^2 values were:

QR: $a_1 = 1.977, a_2 = 0.993, R^2 = 0.998$

Kalman: $a_1 = 1.276, a_2 = -0.304, R^2 = 0.996$

LLS: $a_1 = 0.982, a_2 = -0.018, R^2 = 0.968$

The QR algorithm recovered both coefficients with less than 1% error from the true values, while the LLS coefficients had large errors and could only achieve a model that stated the next sample would be almost the same as the previous sample because $a_1 \approx 1$ and a_2 is nearly zero. Kalman smoothing without the Gaussian Fit step gave an intermediate result with the model coefficients in error by 36% and 70%. It is concluded that the major benefit of the QR algorithm with coarsely quantized data is the expectation step using the Gaussian Fit approximation.

The noisy unit amplitude sine wave with a signal to noise ratio of 1:1 was used as a second test signal into the quantizer input. Application of AIC indicated a model order of 5 or 6 for the QR algorithm and 6 for Kalman smoothing. The recovery ability of the three methods for $m = 6$ AR models is compared in the right hand panel of Figure 7.2. The QR algorithm and Kalman

smoothing provided good reconstruction of the underlying sine wave and both were superior to the LLS reconstruction which provided less filtering of the noise. The R^2 values were:

QR: $R^2 = 0.981$, Kalman: $R^2 = 0.978$, LLS: $R^2 = 0.764$

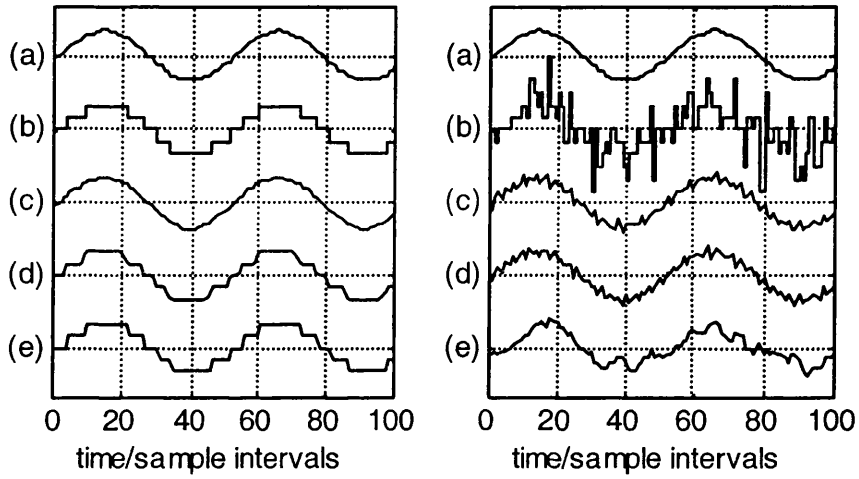


Figure 7-2 *Reconstruction of a sine wave from quantized samples (left) and noisy quantized samples (right). (a): the sine wave signal; (b) the quantized signal; (c) reconstructed with QR; (d) reconstructed with Kalman smoothing; (e) reconstructed with LLS.*

Case 5: Influence of the quantization interval

Figure 7-3 compares the R^2 measure across a range of quantization intervals. The left panel shows the results for the quantized sine wave signal using AR models with model order $m = 2$. The uppermost trend is for the QR algorithm, the lowest for the LLS algorithm and the Kalman smoothing algorithm is in-between. As was suggested by Figure 7-2, the Gaussian Fit element within the QR algorithm gave benefits over and above its Kalman smoothing component. The benefit increased as the quantization interval increased. The right hand panel of Figure 7-3 shows the reconstruction performance of the quantized noisy sine wave using AR models with $m = 6$. The R^2 values for Kalman smoothing and QR were almost identical in this case, and both gave improvements over the LLS method. Thus when the unquantized signal included white noise, no matter what the quantization interval was, the recovery was contributed by Kalman smoothing. It is concluded that the benefit of the Gaussian Fit to the QR algorithm reduces as the influence of quantization reduces relative to the noise.

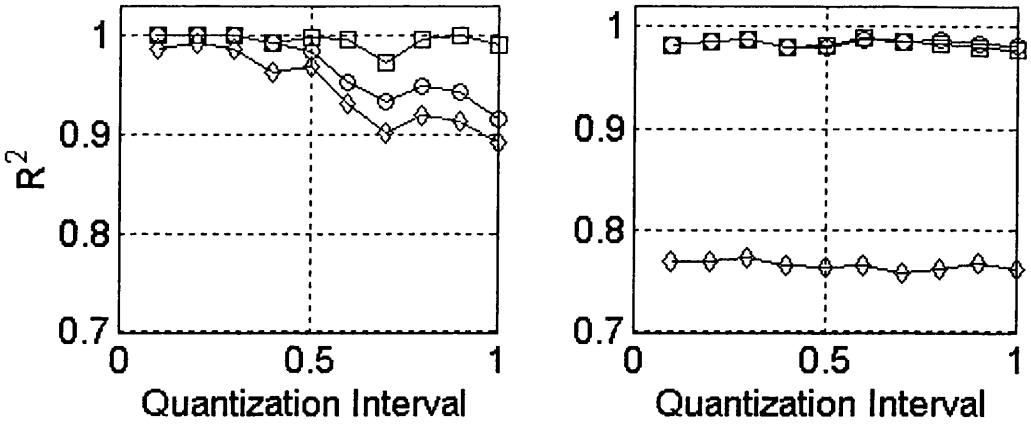


Figure 7-3 *Reconstruction performance for a sine wave from quantized samples (left) and noisy quantized samples (right). QR: squares; Kalman smoothing: circles; LLS diamonds.*

For coarsely quantized signals, the QR (quantized regression) algorithm can recover the underlying signal from quantized observations better than the LLS (linear least squares) algorithm, as measured by the R^2 values. The fundamental reason behind the phenomenon is that QR algorithm based on the Gaussian Fit algorithm and Kalman smoothing can recover some nonlinear parts from the quantized observation, while the LLS algorithm assumes the quantizer has a fixed input-output relationship.

The performance of Kalman smoothing (the same formulation as the QR algorithm but without the Gaussian Fit scheme) was also superior to that of the LLS method. When the effects of quantization were dominant, the performance of the QR algorithm was significantly better than that of Kalman smoothing alone. When the underlying signal was noisy, however, the performances of the QR and Kalman smoothing algorithms were similar. Therefore it is concluded that the Gaussian Fit scheme offers the most improvement when quantization is severe, but that when noise is predominant the majority of the benefit is due to Kalman smoothing.

7.5.2 Experimentation

Figure 7-4 shows the behaviour of the QR, Kalman smoothing and LLS algorithms with pilot plant measurements. The measurements are from the transient response of the pH control in a

buffered fed-batch yeast fermentation process (provided by W.S.Seng and N.Thornhill in the Department of E&E Engineering in University College London). They were quantized by the A/D converter of the pH probe. The aim of the analysis was to recover the underlying smooth transient response of the fermenter pH. The closed loop under proportional-plus-integral control of the process has dominant under-damped complex conjugate poles and therefore the AR time series should have $m = 2$. The s-plane poles are $s = -0.025 \pm 0.108j$ corresponding to damped oscillations with a period of 58 samples and an exponential time constant of 40 samples. These s-plane poles map to z-plane poles at $z = 0.969 \pm 0.105j$.

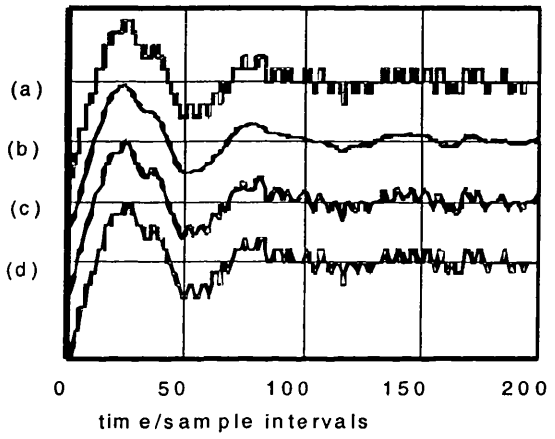


Figure 7-4 Reconstruction of quantized plant data. (a) quantized trend from plant; (b) reconstructed using QR; (c) using Kalman smoothing; (d) using LLS.

Application of AIC to the QR algorithm indicated that the model order should indeed be $m = 2$. The reconstructions of $m = 2$ AR models using QR, Kalman smoothing and the LLS methods are shown in Figure 7-4. It was not possible to determine the R^2 values because the true underlying signal was not known. A quantitative comparison of the z-plane poles shows the QR algorithm has achieved the correct reconstruction because the AR model was:

$$y(n) = 1.914y(n-1) - 0.934y(n-4)$$

which has z-plane poles at $z = 0.957 \pm 0.090j$. These poles are near those of the closed loop process and it is concluded that the AR coefficients determined by the QR algorithm captured the under-damped oscillatory behaviour of the pH response. The other models were:

Kalman: $y(n) = 1.210y(n-1) - 0.23y(n-2)$

LLS: $y(n) = 0.972y(n-1) + 0.008y(n-2)$

both of whose z-plane poles are real. These recovered AR time series have no oscillatory behavior of their own and the reconstructions appear oscillatory only because they are driven by the experimental data.

QR reconstruction of a transient response signal from experimental plant data gave an AR series of the correct order and with z-plane poles close to the true values. It can be concluded that QR reconstruction has the capacity to be useful in the accurate reconstruction of autoregressive time series from quantized process data.

7.6 Conclusions

From the theory, simulation examples and experimentation presented in this chapter, the following conclusions can be reached.

- The Quantized Regression (QR) algorithm itself is robust. The result from the algorithm is insensitive to different initial conditions.
- It is recommended that the data ensemble length be at least 800 samples for consistent estimation when this algorithm is used.
- Akaike Information Criterion (AIC) is a reliable measure to determine model order.
- When the QR algorithm is compared with the LLS (linear Least Square) method and the Kalman smoothing method (the same formulation as the QR algorithm but without the Gaussian Fit scheme), simulation examples and experimentation support the following conclusions:
    - The QR algorithm can recover the underlying signal from quantized observations better than the LLS algorithm, as measured by the R^2 values, no matter when the effects of quantization were dominant or when the underlying signal was noisy. The fundamental reason behind the phenomenon is that QR algorithm based on the Gaussian Fit algorithm and Kalman smoothing can recover some nonlinear parts from the quantized observation, while the LLS algorithm assumes the quantizer has a fixed input-output relationship.

- When the effects of quantization were dominant, the performance of the QR algorithm was significantly better than that of Kalman smoothing alone. When the underlying signal was noisy, however, the performances of the QR and Kalman smoothing algorithms were similar. Therefore it is concluded that the Gaussian Fit scheme offers the most improvement when quantization is severe, but that when noise is predominant the majority of the benefit is due to Kalman smoothing.

7.7 Relevence to the closed-loop identification with quantized data

The aim of the thesis is the identification of the dynamics of a controlled system from closed-loop quantized data. The quantization error is equivalent to an external excitation. Therefore, recovery of the underlying signal from the quantized measurements is an attractive idea. From the descriptions in this chapter, the QR algorithm was implemented successfully.

On the other hand, the QR algorithm needs many iterations because it is based on the EM (Expectation-Maximization) framework, which can guarantee convergency but at a slow convergence rate. This means a lot of computation. This makes the QR algorithm difficult to use on-line. Thus, the QR algorithm was not in the end used in the closed-loop identification procedure.

CHAPTER 8

CONCLUSIONS & RECOMMENDATIONS

This chapter summarises the main conclusions presented in this thesis. The recommendations for future research on this topic are also pointed out.

8.1 Conclusions

8.1.1 Closed-loop identification (CLID) based on quantization

The thesis has shown that it is feasible to identify the dynamics of an open loop system from closed loop data by means of a quantizer inserted into the feedback loop.

The *conclusions* are as follows:

- When quantization interval changes from small to large, the power in the quantization error signal changes from small to large, while the bandwidth of the quantization error on the frequency axes changes from wide to narrow, the nonlinearity of the quantization error also changes from small to large. This provides the basis for the theory of closed loop identification based on quantization.
- When the quantization interval is 'small', closed-loop identification with external excitation (i.e., the two-stage method by Van den Hof and Schrama, 1993 and/or the two-step method by Huang and Shah, 1997) can be used. When the quantization interval is 'large', relay-based identification (i.e. Wang *et al.*, 1997) can be used. Here, 'small' means the quantization interval is in the range from σ_y to $2\sigma_y$ where σ_y is the standard deviation of the process output or, more accurately, from $1.15\sigma_{MV}$ to $1.95\sigma_{MV}$ where σ_{MV} is the standard deviation of the minimum variance for the process output; 'large' means the quantization interval is above $4\sigma_y$.
- In the context of closed-loop identifiability, when the quantization interval is 'small', the quantization error is equivalent to a persistently exciting external excitation. When the quantization interval is 'large', nonlinearity is dominant in the quantiization error. The concept of strong system identifiability requires either one of these conditions to be satisfied. Therefore, closed loop identification based on quantization is enabled in both regimes.

- Simulation examples and experimentation support the new theory. For example, a pilot plant experimental apparatus is used to validate the theory of closed loop identification based on quantization which is presented in Chapter 5 and Chapter 6.
- The insights from simulation examples and experimentation are that for minimum disruption to the closed-loop system, relay-based identification is recommended for "small" noise; otherwise, closed-loop identification with external excitation is recommended for "large" noise. Here, "small" means variance of the process output smaller than 0.1 on a 4 – 20 mA scale; "large" means variance greater than and equal to 0.1 on a 4 – 20 mA scale.

Figure 8-1 summarises the above findings.

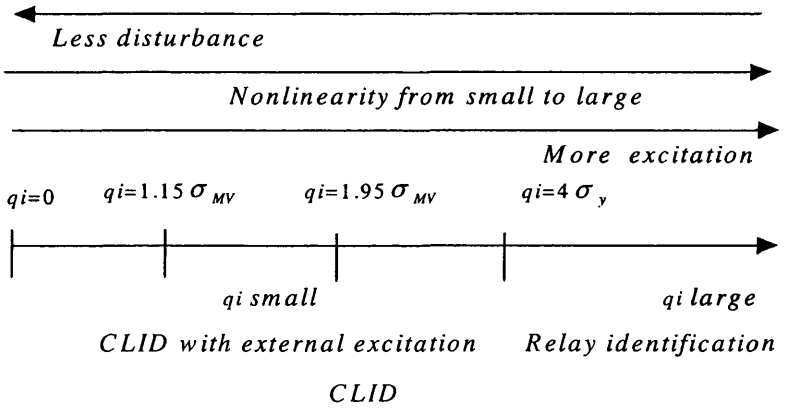


Figure 8-1 *CLID methods suitable for different quantization interval (qi)*

The *benefits* of the proposed CLID scheme based on quantization are as follows:

- Compared with the CLID with external excitation method, there is no requirement for special design of an external excitation. Rather, the quantization interval is adjusted. The quantizer can be implemented by hardware or by software. When the identification is completed, the quantization interval is made negligibly small.
- Compared with the relay-based identification proposed by Åström and Hägglund (1984a and 1984b), this scheme does not need to switch off the controller. Therefore, there is no need for a bumpless transfer scheme and the loop is still under the control of the original controller. Welsh and Goodwin (1999) have previously highlighted this advantage.

The publications closest to this thesis are the papers Goodwin and Welsh (1999) and Welsh and Goodwin (1999). This section draws a comparison and highlights the advances on their work.

- The proposed scheme in this thesis does not require the injection of external excitation. The quantizer provides excitation on both occasions (the quantization interval from σ_y to $2\sigma_y$ and the quantization interval above $4\sigma_y$). By contrast, Welsh and Goodwin (1999) need to inject excitation at a specific frequency to aid the identification.
- The thesis defines the quantizer in a coherent way with three parameters: quantization interval, quantization level and hysteresis width. The two kinds of special quantizers defined in Goodwin and Welsh (1999) and Welsh and Goodwin (1999) are special cases of the new definition.
- The proposed scheme recommends different methods suitable for different quantization intervals. When the quantization interval is small, the quantizer can be used to generate a persistently exciting signal. The two-stage method by Van den Hof and Schrama (1993) or the two-step method by Huang and Shah (1997) can be used directly. When the quantization interval is large, the quantizer acts like a relay. The relay-based identification method by Wang *et al.* (1997) can then be used. In Goodwin and Welsh (1999), a nonparametric DFT method was used for identification, while in Welsh and Goodwin (1999), the traditional frequency response method was used for identification. The benefits of the methods proposed here are that the two-step and/or two-stage method need small signal-to-noise ratio and that the two-step and/or two-stage method and the relay-based identification method are practical.
- The proposed scheme inserts a quantizer in the feedback path, just like a new analogue to digital converter (A/D). Goodwin and Welsh (1999) and Welsh and Goodwin (1999) placed the quantizer after the controller and before the process, just like a digital to analogue converter (D/A). The advantage of the proposed scheme is that it is easier to implement since the quantization interval depends on the analysis of the process output (as presented in the procedure in subsection 5.1.2 and 5.1.3).

8.1.2 Comparison of two categories of CLID methods

The thesis has also compared the advantages and disadvantages by doing case studies (as presented in Chapter 3) on different relay-based identification methods and closed-loop identification with external excitation methods. The main *conclusions* are as follows:

- Closed-loop identification with external excitation

Closed-loop identification with external excitation shows that information of the process over a range of frequencies can be identified and that accurate estimates can be obtained even for small signal-to-noise ratio. It was found that *a priori* information on the critical

frequency of the identified system is required for the design of the persistently exciting signal.

- Relay-based identification

The sustained oscillation required by relay-based identification can be implemented automatically. Therefore, no *a priori* information on the critical frequency of the identified system is needed. The controller must be replaced by a relay, which means disruption to the control system.

8.1.3 Quantized signal recovery

Chapter 7 of the thesis is concerned with quantized signal recovery. Quantized signal recovery is the determination of the true continuous signal from the quantized signal. It is relevant and important because the quantization error is used to substitute an external excitation.

The Quantization Regression (QR) algorithm is based on Ziskand and Hertz (1993). It is a combination of the 'Gaussian Fit' scheme of Curry (1970) with the expectation-maximization (EM) algorithm of Dempster *et al* (1977). From the theory, simulation examples and experimentation presented in Chapter 7, the *conclusions* are as follows:

- The Quantized Regression (QR) algorithm is effective in quantized signal recovery. It was found that the algorithm is robust in recovering quantized signal and is insensitive to different initial conditions. The data ensemble length should be at least 800 samples for consistent estimation when this algorithm is used (the details are presented in p190 of this thesis). The model order may be reliably determined from the Akaike Information Criterion (AIC).
- The QR algorithm is compared with the Linear Least Square (LLS) method and the Kalman smoothing method (the same formulation as the QR algorithm but without Gaussian Fit scheme) in the performance of reconstruction from the quantized signal.

  The QR algorithm can recover the underlying signal from quantized observations better than the LLS (linear least squares) algorithm. The fundamental reason behind the phenomenon is that the QR algorithm, which is based on the Gaussian Fit algorithm and Kalman smoothing, can recover some nonlinear parts from the quantized observation, while the LLS algorithm assumes the quantizer has a fixed input-output relationship.

  When the effects of quantization were dominant, the performance of the QR algorithm was significantly better than that of Kalman smoothing method (i.e. QR without the Gaussian Fit scheme). When the underlying signal was noisy, however, the performances of the QR and Kalman smoothing method were similar. Therefore it is

concluded that the Gaussian Fit scheme offers the most improvement when quantization is severe, but that when noise is predominant the majority of the benefit is due to Kalman smoothing.

8.2 Future research directions

There are many open issues in the use of a quantizer for identification. Three possible developments are discussed below and the potential advantages are explained.

8.2.1 Sampling rate change

In this thesis, the influence of the quantization interval in the inserted quantizer has been studied extensively. The influence of the sampling rate in the inserted quantizer also needs to be explored. The way is to use a different sampling rate for the inserted quantizer, i.e. if the sampling rate for the quantizer input is T_s, the sampling rate for the quantizer output could be $2T_s$, $3T_s$ or any other integer multiple of T_s. It has been observed empirically that changing the sampling rate alters the bandwidth in the frequency domain. The benefit of changing the sampling rate is that it may provide another design parameter in the quest for a persistently exciting signal required for the closed loop identification. Therefore, adjustment of the sampling rate could provide a way forward for closed loop identification.

8.2.2 Different method for identification

The hot topic in the identification literature (Van den Hof and Schrama, 1995) is joint identification and control. The dual Youla parameterization method, which belongs to the indirect approach in the context of closed loop identification, could be a better choice than the two-stage or two-step method currently used. The main advantage of the dual Youla parameterization method is that the obtained estimate of the process is guaranteed to be stabilized by the controller, which clearly is an attractive feature (Forssell and Ljung, 1999).

8.2.3 The QR algorithm

As explained in subsection 8.1.3, quantized signal recovery is relevant to the theme of this thesis. The QR algorithm was implemented successfully but was not in the end used in the closed-loop identification procedure because its large computational requirements make it difficult to use on-line. At present, the quantization error used for closed loop identification is defined to be the true unquantized signal minus the quantized signal. In the future, attempts should be made to speed up the QR algorithm to make it acceptable to use on-line. Thus, the recovered signal from the QR algorithm minus the quantized measurements forms the quantization error, which can be used for closed loop identification.

Appendix A

The QR (Quantized Regression) algorithm

% The purpose of the QR algorithm is to find the underlying continuous signal from the
% quantized measurements. The initialization conditions are given at first. Then the
% algorithm enters a loop for iterative calculation. Inside the loop, the first step is named
% E-step because the conditional expectation $E(x(n) \mid z_n)$ is calculated. The second step is
% named M-step because the complete-data log likelihood is maximized. At the end of the
% loop, the stopping condition is judged to determine whether the program should exit the
% loop.

A.1 Initialization

$$r = 0 \quad s_0^0,\ P_0^0,\ \phi(0),\ Q(0),\ R(0)$$

% $r = 0$ means the algorithm is at the start. Initial conditions include $\mu = s_0^0$, the mean of
% the initial state vector and $\sum_0 = P_0^0$, the covariance of initial state estimate, $\phi(0)$ the
% transition matrix in which elements in the last row are the autoregressive coefficients
% (refer to p178), $Q(0)$ the variance of the model noise (refer to formula 7.2) and
% $R(0)$ the variance of the observation noise

A.2 E-step

% This step is named E-step because the conditional expectation $E(x(n) \mid z_n)$ is calculated,
% in which a modified Kalman smoothing algorithm finds smoothed estimates and their
% covariance matrix $s_n^N, P_n^N, P_{n,n-1}^N$. This step includes Prediction and Filtering,
% Smoothing, Lag-one covariance.

% Prediction and Filtering

For $n = 1,2,\ldots,N$

$$s_n^{n-1} = \phi(r) s_{n-1}^{n-1}$$

% the calculation of the one-step predictor s_n^{n-1} of the state vector by the
% forward recursions

$$P_n^{n-1} = \phi(r) P_{n-1}^{n-1} \phi^T(r) + Q(r)$$

% the conditional error covariance P_n^{n-1} of the state vector

$$\sigma_x^2 = hP_n^{n-1}h^T + \sigma_0^2$$

$$m_x = hs_n^{n-1}$$

% The 'Gaussian Fit' algorithm was proposed in Curry (1970) for a

% discrete-time nonlinear filter that recursively fits a Gaussian distribution

% to the first two moments of the conditional distribution of a state

% vector.Evaluate with 'Gaussian Fit' scheme

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)dt$$

% *erf(X)* is the error function for each element of *X*

$$P(z \in (a,b)) = \frac{1}{2}[erf(\frac{b-\bar{x}}{2^{1/2}\sigma_x}) - erf(\frac{a-\bar{x}}{2^{1/2}\sigma_x})]$$

% the scaling factor which means Prob(x(n)∈A|z_{n-1}), the conditional

% probability density function of x(n) given z_{n-1}

$$E(x(n) \mid z_n) = \bar{x} + \frac{\sigma_x}{P(z \in (a,b))}[\frac{\exp\{-\frac{1}{2}((a-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}} - \frac{\exp\{-\frac{1}{2}(b-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}}]$$

% E(x(n)|Z_n) is the conditional mean

$$\mathrm{cov}(x(n) \mid z_n) = \sigma_x^2[1 + (\frac{a-\bar{x}}{\sigma_x})\frac{\exp\{-\frac{1}{2}((a-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}P(z \in (a,b))} - (\frac{b-\bar{x}}{\sigma_x})\frac{\exp\{-\frac{1}{2}((b-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}P(z \in (a,b))}$$

$$-\frac{1}{P(z \in (a,b))^2}(\frac{\exp\{-\frac{1}{2}((a-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}} - \frac{\exp\{-\frac{1}{2}((b-\bar{x})/\sigma_x)^2\}}{(2\pi)^{1/2}})^2]$$

% cov(x(n)| Z_n) is the conditional covariance

$k_n = P_n^{n-1}h^T(\sigma_x^2)^{-1}$ % the Kalman gain

$s_n^n = s_n^{n-1} + k_n(E(x(n) \mid z_n) - m_x)$

$\tilde{P}_n^n = P_n^{n-1} - k_n h P_n^{n-1}$

$P_n^n = \tilde{P}_n^n + k_n \mathrm{cov}(x(n) \mid z_n)k_n^T$

end

% Smoothing

For n=N,N-1,...,1

$$J_{n-1} = P_{n-1}^{n-1}\phi^T(r)(P_n^{n-1})^{-1} \qquad \%$$

$$s_{n-1}^N = s_{n-1}^{n-1} + J_{n-1}(s_n^N - \phi(r)s_{n-1}^{n-1}) \qquad \% \text{ the smoothed condition mean}$$

$$P_{n-1}^N = P_{n-1}^{n-1} + J_{n-1}(P_n^N - P_n^{n-1})J_{n-1}^T$$

% the smoothed conditional covariance

end

% Lag-one covariance

$$P_{N,N-1}^N = (I - k_N h)\phi(r)P_{N-1}^{N-1}$$

For n=N,N-1,...,2

$$P_{n-1,n-2}^N = P_{n-1}^{n-1}J_{n-2}^T + J_{n-1}(P_{n,n-1}^N - \phi(r)P_{n-1}^{n-1})J_{n-2}^T$$

end

A.3 M-step

% This step is named M-step because the complete-data log likelihood is maximized. The

% likelihood function is maximised by iterative adjustment of the AR coefficients

% $a_1, a_2, ..., a_m$, the estimated signal.

$$B = \sum_{n=1}^{N}\{P_{n,n-1}^N + s_n^N \cdot (s_{n-1}^N)^T\}$$

$$C = \sum_{n=1}^{N}\{P_n^N + s_n^N \cdot (s_n^N)^T\}$$

$$D = \sum_{n=1}^{N}\{P_{n-1}^N + s_{n-1}^N \cdot (s_{n-1}^N)^T\}$$

$$\phi(r+1) = BD^{-1}$$

$$Q(r+1) = \frac{1}{N}(C - BD^{-1}B^T)$$

% $Q(r+1)$ updating of the variance of the model noise (refer to formula 7.2)

$$R(r+1) = \frac{1}{N}\sum_{n=1}^{N}\{(z(n) - hs_n^N)(z(n) - hs_n^N)^T + hP_n^N h^T\}$$

% $R(r+1)$ updating of the variance of the observation noise (refer to formula 7.3)

%Evaluate complete-data log likelihood

$$\ln(\theta \mid \theta_i) = -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}tr\{\Sigma^{-1}[P_0^N + (x_0^N - \mu)(x_0^N - \mu)^T]\}$$

$$-\frac{N}{2}\ln q - \frac{1}{2q}[C_{(m,m)} - B_{(m,:)}\phi_{(m,:)}^T - \phi_{(m,:)}B_{(m,:)}^T + \phi_{(m,:)}D\phi_{(m,:)}^T]\}$$

$$-\frac{N}{2}\ln R - \frac{1}{2R}\sum_{n=1}^{N}[(y_n - hx_n^N)(y_n - hx_n^N)^T + hP_n^N h^T]$$

%initial state vector updating

$s_0^0(r+1) = s_0^N(r)$

A.4 Stopping condition

$r = r + 1$ and judge whether to exit the loop.

% Here the criterion to judge whether to exit the loop is that the innovations form of

% Log Likelihood function (refer to p184) should be stable (for example, smaller

% than 0.01).

Appendix B

Uncertainty Analysis for 'Low-order Modeling from Relay Feedback (Wang *et al.*, 1997a)'

The aim of this appendix is to provide the uncertainty analysis for the results of the relay-based identification method 'Low-order Modeling from Relay Feedback (Wang *et al.*, 1997a)', which was duplicated in subsection 3.3.3 of Chapter 3 as a case study. In this appendix, the theory for uncertainty analysis will be cited first. The theory for the relay-based identification method (Wang *et al.*, 1997a) is then briefly reviewed. Uncertainty analysis for simulation results then follows. Conclusion is given at the end.

B.1 Theory for uncertainty analysis

The following paragraph is cited from the textbook by Taylor (1997). It will be used for analysis in this appendix.

> Uncertainty in a function of several variables
>
> Suppose that $x,\ldots,z$ are measured with uncertainties $\delta_x,\ldots,\delta_z$ and the measured values are used to compute the function $q(x,\ldots,z)$. If the uncertainties in $x,\ldots,z$ are independent and random, then the uncertainty in q is
>
> $$\delta q = \sqrt{(\frac{\partial q}{\partial x}\delta x)^2 + \ldots + (\frac{\partial q}{\partial z}\delta z)^2} \qquad \text{(B-1)}$$
>
> (Taylor, 1997:75)

B.2 The theory for the relay-based identification method (Wang et al., *1997a)*

The theory for the relay-based identification method 'Low-order Modeling from Relay Feedback (Wang *et al.*, 1997a)' is presented in Chapter 3 (subsection 3.3.1). The most important formulas are written here again.

$$K = G(0) = \frac{\int_0^{T_{u1}+T_{u2}} y(t)dt}{\int_0^{T_{u1}+T_{u2}} u(t)dt} \quad \text{(B-2)}$$

$$\theta = \ln \frac{(\mu_0 + \mu)K - \varepsilon}{(\mu_0 + \mu)K - A_\mu} \quad \text{(B-3)}$$

or

$$\theta = \ln \frac{(\mu - \mu_0)K - \varepsilon}{(\mu - \mu_0)K + A_d} \quad \text{(B-4)}$$

$$T = T_{u1}(\ln \frac{2\mu K e^{\theta} + \mu_0 K - \mu K + \varepsilon}{\mu K + \mu_0 K - \varepsilon})^{-1} \quad \text{(B-5)}$$

or

$$T = T_{u2}(\ln \frac{2\mu K e^{\theta} - \mu_0 K - \mu K + \varepsilon}{\mu K - \mu_0 K - \varepsilon} \quad \text{(B-6)}$$

$$L = T\theta \quad \text{(B-7)}$$

B.3 Uncertainty analysis

From the simulation time trend presented Figure 3-18 (which is reproduced here for convenience), the following measurements can be obtained from one cycle.

$$A_u = 0.317 \pm 0.001 \qquad A_d = -0.208 \pm 0.001$$

$$T_{u1} = 4.291 \pm 0.001 \qquad T_{u2} = 7.278 \pm 0.001$$

Using the method of Wang *et al* (1997), the best estimated model is

$$K = 1.007; \qquad T = 10.084; \qquad L = 1.998$$

Therefore, the estimated transfer function model is

$$\hat{G} = \frac{1.007e^{-1.998s}}{10.084s + 1} \quad \text{(B-8)}$$

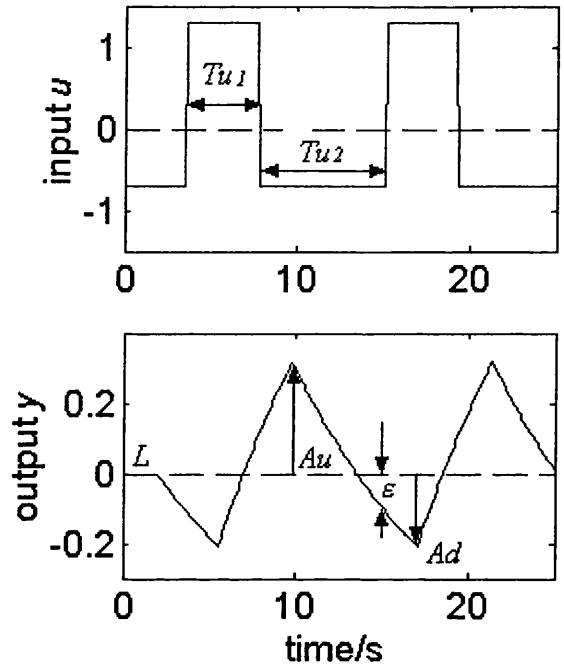


Figure 3-18 *Time trends by using the TD1 method*

The fractional uncertainties are

$$\frac{\delta A_u}{A_u} = \frac{0.001}{0.317} = 0.32\% \qquad \frac{\delta A_d}{A_d} = \frac{0.001}{0.208} = 0.48\%$$

$$\frac{\delta T_{u1}}{T_{u1}} = \frac{0.001}{4.291} = 0.0233\% \qquad \frac{\delta T_{u2}}{T_{u2}} = \frac{0.001}{7.278} = 0.0137\%$$

The uncertainty analysis includes four steps as described below.

Step 1: the uncertainty in K

The gain K is a derived variable, its uncertainty cannot be read directly from the simulation, but from repeated calculation with formula (B-2).

$$K = 1.007 \pm 0.003; \qquad \frac{\delta K}{K} = \frac{0.003}{1.007} = 0.3\% \qquad \text{(B-9)}$$

Step 2: the uncertainty in θ

θ can be calculated from formula (B-3) or (B-4). Here (B-3) is chosen. The uncertainties in K and A_u will contribute to the uncertainty in θ.

$$\delta\theta_K = \left|\frac{\partial\theta}{\partial K}\right|\delta K = \left|\frac{(\mu_0+\mu)K - A_u}{(\mu_0+\mu)K-\varepsilon}\cdot\frac{(\mu_0+\mu)(-A_u+\varepsilon)}{[(\mu_0+\mu)K-A_u]^2}\right|\delta K = 0.0007 \quad \text{(B-10)}$$

$$\delta\theta_{A_u} = \left|\frac{\partial\theta}{\partial A_u}\right|\delta A_\mu = \left|\frac{(\mu_0+\mu)K - A_u}{(\mu_0+\mu)K-\varepsilon}\cdot\frac{(\mu_0+\mu)K-\varepsilon)}{[(\mu_0+\mu)K-A_u]^2}\right|\delta A_u = 0.001 \quad \text{(B-11)}$$

$$\delta\theta = \sqrt{(\delta\theta_K)^2+(\delta\theta_{A_u})^2} = \sqrt{(\frac{\partial\theta}{\partial K}\delta K)^2+(\frac{\partial\theta}{\partial A_u}\delta A_u)^2} = 0.0012 \quad \text{(B-12)}$$

Therefore, $\theta = 0.1999 \pm 0.0012$

Step 3: the uncertainty in T

T can be calculated from formula (B-5) or (B-6). Here (B-5) is chosen. The uncertainties in K, θ and T_{u1} will contribute to the uncertainty in T.

$$\delta T_{Tu1} = \left|\frac{\partial T}{\partial T_{u1}}\right|\delta T_{u1} = \left|(\ln\frac{2\mu Ke^{\vartheta}+\mu_0 K-\mu K+\varepsilon}{\mu K+\mu_0 K-\varepsilon})^{-1}\right|\delta T_{u1} = 0.0023 \quad \text{(B-13)}$$

$$\delta T_K = \left|\frac{\partial T}{\partial K}\right|\delta K = \left|T_{u1}(\ln\frac{2\mu Ke^{\vartheta}+\mu_0 K-\mu K+\varepsilon}{\mu K+\mu_0 K-\varepsilon})^{-2}\cdot\frac{2(\mu e^{\vartheta}+\mu_0)}{(2\mu Ke^{\vartheta}+\mu_0 K-\mu K+\varepsilon)(\mu K+\mu_0 K-\varepsilon)}\right|\delta K$$

$$= 0.0222 \quad \text{(B-14)}$$

$$\delta T_\theta = \left|\frac{\partial T}{\partial\theta}\right|\delta\theta = \left|T_{u1}(\ln\frac{2\mu Ke^{\vartheta}+\mu_0 K-\mu K+\varepsilon}{\mu K+\mu_0 K-\varepsilon})^{-2}\cdot\frac{-2\mu Ke^{\vartheta}}{(2\mu Ke^{\vartheta}+\mu_0 K-\mu K+\varepsilon)}\right|\delta\theta = 0.0087$$

(B-15)

$$\delta T = \sqrt{(\delta T_{Tu1})^2+(\delta T_K)^2+(\delta T_\theta)^2} = \sqrt{(\frac{\partial T}{\partial T_{u1}}\delta T_{u1})^2+(\frac{\partial T}{\partial K}\delta K)^2+(\frac{\partial T}{\partial\theta}\delta\theta)^2} = 0.0240$$

(B-16)

$T = 10.084 \pm 0.024$

Step 4: the uncertainty in L

L can be calculated from formula (B-7). The uncertainties in θ and T will contribute to the uncertainty in L.

$$\delta L_T = \left|\frac{\partial L}{\partial T}\right| \delta T = |\theta| \delta T = 0.0002 \tag{B-17}$$

$$\delta L_\theta = \left|\frac{\partial L}{\partial \theta}\right| \delta \theta = |T| \delta \theta = 0.0121 \tag{B-18}$$

$$\delta L = \sqrt{(\delta L_T)^2 + (\delta L_\theta)^2} = \sqrt{(\left|\frac{\delta L}{\partial T}\right| \delta T)^2 + (\left|\frac{\delta L}{\partial \theta}\right| \delta \theta)^2} = 0.0121 \tag{B-19}$$

$$L = 1.998 \pm 0.0121$$

B.4 Conclusion

From section B.3, the derived variables that construct the identified model and their uncertainties are summarized in the following.

$$K = 1.007 \pm 0.003$$

$$T = 10.084 \pm 0.024$$

$$L = 1.998 \pm 0.0121$$

The conclusion is that the identified model is reliable.

REFERENCES

[1] Åström,K.J.(1991), Assessment of achievable performance of simple feedback loops, *Int. J. Adaptive Control and Signal Processing*, 5, 3-19.

[2] Åström,K.J. and Hagglünd,T.(1984a), Automatic tuning of simple regulators with specifications on phase and amplitude margins, *Automatica*, Vol.20, No.5, pp645-651.

[3] Åström,K.J. and Hagglünd,T.(1984b), Automatic tuning of simple regulator, *Proceedings IFAC 9th World Congress*,pp1867-1872.

[4] Åström,K.J., Hägglund,T., Hang,C.C. and Ho,W.K. (1993), Automatic tuning and adaptation for PID controllers - a survey, *Control. Engn. Prac.*, 1,699-714.

[5] Åström,K.J. and Hägglund,T. (1995), *PID controllers*, 2nd ed, International Society for Measurement and Control, Research Triangle Park, N.C.

[6] Åström,K.J., Hang,C.C., Persson,P. and Ho,W.K.(1992), Towards intelligent PID control, *Automatica*, 28, 1-9.

[7] Åström,K.J.and Wittenmark, B. (1997), *Computer-Controlled Systems*, 3rd ed, Prentice Hall, New Jersey, USA.

[8] Atherton, D.P. (1975), *Nonlinear control engineering*, Van Nostrand Reinhold, London.

[9] Bartree,J.F. and McFarlane,R.C.(1998), Identification of linear systems operating in closed loop, *Proceedings of IChemE: Advances in Process ontrol*, Swansea, 2-3 Sept. 1998.

[10] Besancon-Voda,A and Roux-Buisson,H (1997), Another version of the relay feedback experiment, *J.Proc.Cont.*,7,303-308.

[11] Bezergianni,S. and Georgakis,C.(2000), Controller performance assessment based on minimum and open-loop output variance, *Control Engineering Practice*, 8, 791-797.

[12] Bi,Q.,Wang,Q.G. and Hang,C.C. (1997), Relay-based estimation of multiple points on process frequency response, *Automatica*, 33, 1753-1757.

[13] Callafon,R., *Feedback oriented identification for enhanced and robust control*, Ph.D. Thesis, Delft University of Tech., The Netherlands, Sept, 1998.

[14] Cavanaugh, J.E.C. and Shumway, R.H.(1998), An Akaike information criterion for model selection in the presence of incomplete data, *Journal of Statistical Planning and Inference*, Vol.67, pp45-65.

[15] Chang,R.C., Shen,S.H., Yu,C.C. (1992), Derivation of transfer function from relay feedback systems, *Ind. Eng. Chem. Res.*, 31, 855-860.

[16] Chatfield, C. (1989), *Time series analysis: an introduction*, London, Chapman and Hall Ltd.

[17] Crowe, J. and Johnson, M.A. (2000), Open and closed loop identification by a phase locked loop identifier module, *ADCHEM 2000*, 761-766.

[18] Curry, R. E. (1970), *Estimation and control with quantized measurements*, Cambridge, MA: MIT Press.

[19] Dempster,A.P., Laird,N.M. and Rubin,D.B.(1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistic Society*, Vol.B-39, pp1-37.

[20] Desborough,L. and Harris,T.J.(1992), Performance assessment measures for univariate feedback control, *Can.J.Chem.Eng.*, 70, 1186-1197.

[21] Esmaili, A., J.F. MacGregor, and P.A. Taylor, (2000). Direct and two-step methods for closed-loop identification: a comparison of asymptotic and finite data set performance, *J. Proc. Cont.*, 10, 525-537.

[22] Forssell,U. and Ljung,L.(1998), A projection method for closed-loop identification, Internal Report no. LiTH-ISY-R-2080, Automatic Control Group, Linkoping University, Sweden.

[23] Forssell,U and Chou,C.T.(1998), Efficiency of prediction error and instrumental variable methods for closed-loop identification, Internal Report no. LiTH-ISY-R-2015, Automatic Control Group, Linkoping University, Sweden.

[24] Forssell,U. and Ljung,L.(1999), closed-loop identification revisited, *Automatica*, 35, 1215-1241.

[25] Fox,P.D. and Godfrey, K.R.(1998), Autotuning with multiharmonic perturbations - Part 1 nonparametric closed loop identification, *UKACC International Conference on CONTROL '98*, 1-4 September.

[26] Friman,M. and Waller,K.V. (1997), A two-channel relay for auto-tuning, *Ind. Eng. Chem. Res.*, 36, 2662-2671.

[26] Gevers, M., Ljung L. and Van den Hof, P (2001), Asymptotic variance expressions for closed-loop identification, *Automatica*, Vol. 37, No. 5, 781-786.

[27] Goodwin, G.C. and Welsh, J.S. (1999), Analysis of a novel method of autotuning a multivariable plant based on quantization, *Proceedings of American Control Conference*, San Diego, California.

[28] Gopaluni, R.B., Patwardhan, R.S. and Shah, S.L. (2002a), Bias distribution in MPC relevant identification, *Proceedings of 15th IFAC World Congress,*Barcelona, Spain.

[29] Gopaluni, R.B., Patwardhan, R.S. and Shah, S.L. (2002b), Experimental design for MPC relevant identification, *Proceedings of American Control Conference,* Ancrage, Alaska, USA.

[30] Gustavsson,I., Ljung,L. and Soderstrom,T.(1977), Identification of processes in closed loop - identifiability and accuracy aspects, *Automatica*, 13, 59-75.

[31] Hakvoort,R.G., Schrama,R. and Van Den Hof,P.M.J.(1994), Approximate identification with closed-loop performance criterion and application to LQG feedback design, *Automatica*, 30, 679-690.

[32] Hang,C.C., Åström,K.J. and Ho,W.K. (1993), Relay auto-tuning in the presence of static load disturbance, *Automatica*, 29, 563-564.

[33] Hang,C.C., Åström,K.J. and Wang, Q.G. (2002), Relay feedback auto-tuning of process controllers —a tutorial review, *J. Proc. Cont.*, 12, 143-162.

[34] Harris, T.J.(1989), Assessment of control loop performance, *Can.J.Chem.Eng.*, 67, 856-861.

[35] Harris,T., Seppala,C.T. and Desborough,L.D. (1999), A review of performance monitoring and assessment techniques for univariate and multivariate control systems, *J. Proc. Cont.*, 9, 1-17.

[36] Hjalmarsson,H., Gevers,M. and Bruyne,F.(1996), For model-based control design, closed-loop identification gives better performance, *Automatica*, 32, 1659-1673.

[37] Huang,B., *Multivariate statistical methods for control loop performance assessment*, Ph.D. Thesis, University of Alberta, Canada,1997.

[38] Huang,B. and Shah,S.L.(1997), closed-loop identification: a two step approach, *J. Proc. Cont.*, 7, 425-438.

[39] Jamshidian,M. and Jennrich,R.(1997), Acceleration of the EM algorithm by using Quasi-Newton methods, *J. R. Statist. Soc.* B, Vol.59, No.3, pp569-587.

[40] Jofriet,P., Seppala,C., Harvy,M., Surgenor,B. and Harris,T. (1996), An expert system for control loop performance, *Pulp & Paper Canada*, 97,70-73.

[41] Kantz, H and Schreider T (1995), *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, UK.

[42] Kendra,S.J. and Cinar,A.(1997), Controller performance assessment by frequency domain techniques, *J. Proc. Cont.*, 7, 181-194.

[43] Ko,B. and Edgar,T.F (1998)., Assessment of achievable PI control performance for linear processes with dead time, *Proceedings of American Control Conference*, Philadelphia, PA.

[44] Krauss, T.P., Shure L. and Little J.N. (1994), Signal Processing Toolbox User's Guide, *The Mathworks Inc.*, USA.

[45] Kumar,V.P. and Rangaiah,G.P. (1999), Process Identification from Closed-loop Response by Global Optimization, AIChE Spring Meeting.

[46] Leonard,F. and Oubrahim,R. (1998), Two steps auto-tuning in the presence of static load disturbance, *UKACC International Conference on CONTROL '98*,1-4 September, 1998, 1339-1343.

[47] Li,W., Eskinat,E. and Luyben,W.L. (1991), An Improved Auto-tune Identification Method, *Ind. Eng. Chem. Res.*, 30, 1530-1541.

[48] Little,R. and Rubin,D.B.(1986), Statistical Analysis with Missing Data, NY,USA., John Wiley and Sons.

[49] Ljung,L.(1987), *System Identification: theory for the user*, Englewood Cliffs, NJ: Prentice-Hall.

[50] Ljung, L.(1991), *System Identification Toolbox User's Guide*, The Mathworks Inc., USA.

[51] Ljung.L and Forssell,U (1997), Variance Results for Closed-Loop Identification Methods, *Proceedings of the 36th Conference on Decision & Control*, san Diego, Califoprnia, USA, 2435-2440.

[52] Ljung,L. and Forssell,U.(1999), An alternative Motivation for the Indirect Approach to Closed-Loop ID, *IEEE Trans. Automat Control.*, 44, 2206-2209.

[53] Luo,R., Qin,S.J. and Chen,D. (1998), A New Approach to Closed-Loop Auto-tuning for Proportional-Integral-Derivative Controller, *Ind. Eng. Chem. Res.*, 37, 2462-2468.

[54] Luyben,W.L. (1987), Derivation of Transfer Function for Highly Nonlinear Distillation Columns, *Ind. Eng. Chem. Res.*, 26, 2490-2495.

[55] Lynch,C.B. and Dumont,G.A.(1996), Control Loop Performance Monitoring, *IEEE Trans. on Control Systems Technology*, 4, 185-192.

[56] MacGregor,J.F. and Fogal,D.T. (1995), Closed-loop identification: the role of the noise model and pre-filers, *J. Proc. Cont.*, 5, 163-171.

[57] Meng,Xiao-li and Rubin,D.B.(1991), Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *Journal of the American Statistical Association*, Vol.86, No416, pp899-909.

[58] Nikolaou,M and Eker,S.A. (1998), Adaptive Control through On-line Optimization: the MPCI Paradigm and Variants, *AIChE Spring Conference.*

[59] Norton, J.P. (1986), *An introduction to identification*, Academic Press, London, UK.

Park,J.H., Sung,S.W. and Lee, I.B. (1997), Improved Relay auto-tuning with static load disturbance, *Automatica*, 33, 711-715.

[60] Proakis, J. and Manolakis, D. (1988), *Introduction to Digital Signal Processing*, London, Collier Macmillan, 1988.

[61] Qin,S.J.(1998), Control Performance Monitoring - A Review and Assessment, *Computer and Chemical Engineering*, 23, 173-186.

[62] Ramirez, R.W.(1985), *The FFT Fundamentals and Concepts*, Prentice Hall, Englewood Cliffs, NJ.

[63] Scheaffer,R.L. and Mcclave,J.T.(1995), Probability and Statistics for Engineers, Belmont, Duxbury Press.

[64] Schei,T.S. (1992), A Method for Closed Loop Automatic Tuning of PID Controller, *Automatica*, 28, 587-591.

[65] Schrama, R.J.P.(1991), An Open-Loop Solution to the Approximate Closed-Loop Identification Problem, *IFAC Identification and System Parameter Estimation*, Budapest, Hungary, 1991.

[66] Shen,S.H., Yu,H.D. and Yu,C.C. (1996**a**), Use of Saturation-Relay feedback for Auto-tune Identification, *Chem. Eng. Sci.*, 51, 1187-1198.

[67] Shen,S.H., Wu,J.S. and Yu,C.C. (1996**b**), Use of Biased-Relay Feedback for System Identification, *AIChE Journal*, 42, 1174-1180.

[68] Shen,S.H., Wu,J.S. and Yu,C.C. (1996**c**), Auto-tune Identification under Load disturbance, *Ind. Eng. Chem. Res.*, 35, 1642-1651.

[69] Shumway,R.H and Stoffer,D.S.(1982)., An approach to time series smoothing and forecasting using the EM algorithm, *Journal of Time Series Analysis*, Vol.3, No.4, pp253-263.

[70] Shumway,R.H and Stoffer,D.S.(2000)., Time series analysis and its applications, New York. London. Springer.

[71] Söderström,T., Stoica,P. and Trulsson,E.(1987), Instrumental variable methods for closed-loop systems, In *Proc. 10th IFAC World Congress*, 364-369.

[72] Söderström, T. and Stoica, P. (1989), *System Identification*, Prentice Hall International, UK 1989.

[73] Stanfelj,N., Marlin T. and MacGregor,J (1993)., Monitoring and Diagnosing Process Control Performance: The Single Loop Case, *Ind. Eng. Chem. Res.*, 32, 301-314.

[74] Stephanopoulos,G.(c1984), *Chemical Process Control: An Introduction to Theory and Practice,* Englewood Cliffs, Prentice-Hall.

[75] Sun,L., Ohmori,H. and Sano,A.(2001), Output intersampling approach to direct closed-loop identification, *IEEE Transaction on Automatic Control*, Vol. 46, No. 12, 1936-1941.

[76] Sung,S.W., Park,J.H. and Lee,I.B. (1995), Modified Relay Feedback Method, *Ind. Eng. Chem. Res.*, 34, 4133-4135.

[77] Taylor J.R.(1997) , An Introduction to Error Analysis, 2nd Edition, University Science Books, Sausalito, California.

[78] Thornhill,N.F., Oettinger,M., Fedenczuk,P. (1999), Refinery-wide control loop performance assessment, *J. Proc. Cont.*, 9, 109-124.

[79] Thornhill, N.F., Vinter, R. and Coclough, M.(2000), *Closed-loop identification using quantized data*, Project proposal to EPSRC GR/R06687.

[80] Thornhill, N.F., Shah, S.L., and Huang, B., 2001, Detection of distributed oscillations and root-cause diagnosis, *CHEMFAS4*, Korea, 7-8 Jun, 167-172.

[81] Van den Hof,P.M.J. and Schrama,R.(1993), An Indirect Method for Transfer Function Estimation from Closed Loop Data, *Automatica*, 29, 1523-1527.

[82] Van Den Hof,P.M.J. and Schrama,R.(1995), Identification and Control - Closed-Loop Issues, *Automatica*, 31, 1751-1770.

[83] Wang,Q.G., Hang,C.C. and Zou,B (1997a), Low-Order Modeling from relay Feedback, *Ind. Eng. Chem. Res.*, 36, 375-381.

[84] Wang,Q.G., Hang,C.C and Bi,Q.(1997b), Process Frequency Response Estimation from relay feedback, *Control. Engn. Prac.*, 5,1293-1302.

[85] Wang,Q.G., Hang,C.C., Zhu,S.A. and Bi,Q. (1999), Implementation and testing of an advanced relay auto-tuner, *J. Proc. Cont.*, 5, 163-171.

[86] Welsh, J.S. and Goodwin, G.C. (1999), A novel mechanism for autotuning based on quantization, *Proceedings IFAC 14th World Congress*, Beijing China.

[87] Yuwana,M. and Seborg,D.E. (1982), A New Method for On-line Controller Tuning, *AIChE Journal*, 28, 434-439.

[88] Zhang,Y.,Wen,C. and Soh,Y. C. (1997), Indirect Closed-Loop Identification by Optimal Instrumental Variable Method, *Automatica*, 33, 2269-2271.

[89] Zheng,W.X. and Feng,C.(1995),A Bias-correction Method for Indirect identification of Closed-loop Systems, *Automatica*, 31, 1019-1034.

[90] Zhu, Y. and Butoyi, F. (2002), Case studies on closed-loop identification for MPC, *Control Engineering Practice*, 10, 403-417.

[91] Ziskand, I. &Hertz, D.(1993), Multiple frequencies and AR parameters estimation from one bit quantized signal via the EM algorithm, *IEEE Trans Signal Processing*, 41, 3202-3206.