VARIABLE REGRESSION ESTIMATION OF UNKNOWN SYSTEM DELAY

By

Ashraf Elnaggar

B. Sc. Electrical Engineering, Cairo University, Egypt, 1982
M. Sc. Electrical Engineering, Cairo University, Egypt, 1985

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

in
THE FACULTY OF GRADUATE STUDIES
ELECTRICAL ENGINEERING

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
December 1990

© Ashraf Elnaggar, 1990
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Electrical Engineering

The University of British Columbia
Vancouver, Canada

Date Feb. 5, 1991
Abstract

This thesis describes a novel approach to model and estimate systems of unknown delay. The a-priori knowledge available about the systems is fully utilized so that the number of parameters to be estimated equals the number of unknowns in the systems. Existing methods represent the single unknown system delay by a large number of unknown parameters in the system model.

The purpose of this thesis is to develop new methods of modelling the systems so that the unknowns are directly estimated. The Variable Regression Estimation technique is developed to provide direct delay estimation. The delay estimation requires minimum excitation and is robust, bounded, and it converges to the true value for first-order and second-order systems. The delay estimation provides a good model approximation for high-order systems and the model is always stable and matches the frequency response of the system at any given frequency. The new delay estimation method is coupled with the Pole Placement, Dahlin and the Generalized Predictive Controller (GPC) design and adaptive versions of these controllers result. The new adaptive GPC has the same closed-loop performance for different values of system delay. This was not achievable in the original adaptive GPC. The adaptive controllers with direct delay estimation can regulate systems with dominant time delay with minimum parameters in the controller and the system model. The delay does not lose identifiability in closed-loop estimation. Experiments on the delay estimation show excellent agreement with the theoretical analysis of the proposed methods.
Acknowledgment

I would like to express my deep thanks and gratitude to my research supervisor Dr. G. A. Dumont for his guidance and supervision throughout my study. I'd like also to thank him and the Pulp And Paper Research Institute of Canada (PAPRICAN) for providing the financial support and computer facilities since the beginning of my study. I'd like also to thank Dr. W. G. Dunford who accepted me first in UBC and provided some financial assistance.

The basic idea of Variable Regression Estimation was born when I was taking EE 570 course and I thank Dr. M. S. Davies, the course instructor, for referring to the problem of systems of unknown delay in his course and for encouraging me to study this problem in depth.

Every idea, problem and result in my thesis was discussed with my colleague Abdel-Latif Elshafei and many improvements of the basic VRE are the results of our continuing discussions. He has also read the different versions of the thesis and provided me with valuable suggestions.

At the end of my student life, I thank all who have helped me through it. Special thanks and gratitude to my late father and to my mother who have provided me with good life and excellent studying environment for many years and to my brother Ayman who have always believed in me. I also thank my teachers at the different schools and universities in Egypt who encouraged and helped me.

To my little girl Miriam I give my thanks for the blessing and joy that came to my life with her and for making me believe more in adaptive control. Finally, my deep gratitude and thanks to my wife Mona for the unlimited support and encouragement.
she gives me and for the lovely home life that we share.
# Table of Contents

Abstract  

Acknowledgment  

List of Figures  

1 Introduction  

1.1 Background  

1.2 Motivation for the Research Topic  

1.3 Main Contributions of the Thesis  

1.4 Thesis Outline  

2 Introduction to Delay Estimation  

2.1 Methods of Delay Estimation  

2.1.1 The Extended Numerator Polynomial Method  

2.1.2 The Rational Approximation Method  

2.1.3 The Cross-Correlation Method  

2.2 Conclusions  

3 Delay Estimation Using Variable Regression  

3.1 Variable Regression  

3.2 The Estimated Model Variable Regression Estimation (EMVRE)  

3.2.1 Advantages and disadvantages of the EMVRE  

3.3 Conclusions  

4 Conclusions  

5 Appendices  

6 References  

7 Index  

v
7 Conclusions

7.1 Summary of the Work ............................................. 114
7.2 Future Research Directions .................................... 116

Bibliography ............................................................. 117
List of Figures

3.1 Minimization of $J$ ................................................. 21
3.2 EMVRE of a first order system................................. 27
3.3 Estimation and adaptive P-P control using the EMVRE... 28
3.4 EMVRE at different sampling rates............................ 30
3.5 EMVRE of a high order system................................. 31

4.6 Impulse response of $G_1$ .......................................... 37
4.7 Delay estimation using Ziegler-Nichols and FMVRE...... 41
4.8 FMVRE of a first order system................................. 44
4.9 Minimum $T_s$ required for a second order system...... 46
4.10 FMVRE of a second order system.............................. 48
4.11 Response of a high order system.............................. 49
4.12 FMVRE of a high-order system............................... 51
4.13 FMVRE in presence of coloured noise........................ 54
4.14 FMVRE using only one sinusoid............................... 59
4.15 Bode plot of the third-order system......................... 63
4.16 Delay estimation for the case of model order mismatch 65
4.17 Parameter estimation for the case of model order mismatch 66
4.18 Delay estimation for system with fractional delay........ 70
4.19 Parameter estimation for system with fractional delay... 71

5.20 Structure of the Estimators and the adaptive controller 74
5.21 Open-loop response and FMVRE delay estimation........ 77
5.22 Adaptive PP/FMVRE response and delay estimation ............... 81
5.23 Adaptive PP/FMVRE parameter estimates ......................... 82
5.24 Adaptive PP/extended polynomial response ....................... 83
5.25 Adaptive PP/EMVRE response and delay estimation ............... 84
5.26 Adaptive Dahlin/FMVRE response and delay estimation .......... 86
5.27 Adaptive Dahlin/FMVRE parameter estimates ..................... 87
5.28 Adaptive Dahlin/extended polynomial response .................. 88
5.29 Adaptive Dahlin/EMVRE response and delay estimation .......... 89
5.30 Adaptive GPC/FMVRE response and delay estimation ............ 95
5.31 Adaptive GPC/FMVRE parameter estimates ....................... 96
5.32 Adaptive GPC/extended polynomial response ..................... 97
5.33 Adaptive GPC/EMVRE response and delay estimation ............ 98
5.34 Correlation functions of the process noise ...................... 101

6.35 Temperature control experiment .................................. 105
6.36 Delay estimation and temperature variation ...................... 107
6.37 Schematic representation of a pulp refiner ...................... 109
6.38 Measurements on the pulp refiner ................................ 111
6.39 Delay and gain estimation of the pulp refiner ................... 112
6.40 Correlation functions of the pulp refiner measurements ....... 113
Chapter 1

Introduction

1.1 Background

The field of parameter estimation is of great importance in control applications where it is generally difficult, if not impossible, to develop an exact model of the process to be controlled. Many processes are so complex that developing an exact model from the basic principles is very difficult. Even when an exact model can be derived, the parameters of this model may not be known or may change with time, temperature, operating point, etc. in an unknown way. For example, the resistance of an electric motor depends on its temperature, which is not usually measurable. For these systems the parameter estimation methods offers a simple way of identifying the process parameters using input-output measurements. Parameter estimation is a very active research area in control, signal processing and statistics literature. The introduction and successful use of the adaptive techniques for automatic control and signal processing over the last 30 years has encouraged research in estimation methods and led to important developments of the parameter estimation methods and the analysis of their properties.

The least-squares method is the most important method of parameter estimation and is heavily used in practice. The basic idea can be traced back to Gauss (1809). The recursive form of least-squares has apparently been found by several authors. The original reference seems to be Plackett (1950). There exist many papers dealing with
different aspects of the least-squares method. Studies of the consistency properties and the implementation aspects can be found in many papers and books, for example Ljung and Söderström (1987) and Goodwin and Sin (1984). In this thesis, least-squares estimation methods are applied in new ways to important control problems. Detailed surveys will be given at the time of introducing those problems.

1.2 Motivation for the Research Topic

The standard estimation algorithms have already been applied to thousands of systems. Although many systems can be systematically modelled by a linear input-output model, some a-priori information about the system is always required to limit the time and effort spent in the modelling process. There is an infinite number of ways that describe the input-output relations of the system. For example a first-order Auto-Regressive Moving-Average (ARMA) model can be represented by an equivalent infinite order Moving-Average (MA) model. In this case, the ARMA model has only two unknowns (the pole location and the gain), while the equivalent MA model has an infinite number of unknowns (the impulse response coefficients). There is a direct correlation between the number of unknown parameters in the model and the signal excitation required to estimate those parameters correctly. Also, the rate of convergence of the estimation process decreases rapidly when the number of parameters increases. For the first-order example, it is the a-priori knowledge of the system order that makes it possible to represent the system by only two parameters in the ARMA model instead of thirty or more parameters in the MA model.

It is surprising to discover that many adaptive control applications do not utilize some crucial information available about the systems they control. In this thesis, an
important case of adaptive control is studied and the a-priori available system information is fully utilized. A minimum number of parameters are estimated to represent the system. Accurate, fast and easy estimation of models results from the full use of the available information. The adaptive control area covered in this thesis is:

*Systems with delay.* These systems are very common in industry. Conveyor belts, long pipe-lines or any transportation line are typical sources of system delay where the system input goes to one end of the transportation line (e.g. output valve of a reservoir tank) and the output is measured at the other end of the line (e.g. the level of receiving tank). Traditional on-line methods of delay estimation do not directly estimate the system delay but rather replace it by a large number of unknown parameters in the system model. In this thesis, the unknown delay is directly estimated and no additional parameters are used in the system model.

### 1.3 Main Contributions of the Thesis

A novel method of on-line delay estimation using the new concept of Variable Regression Estimation (VRE) is developed and demonstrated analytically, on computer simulations and experimentally.

- The variable regression can be added to any standard parameter estimation with minimum additional computations. The number of estimated parameters is the same as that in the known delay case. This means faster convergence and less excitation required. In the case of known process parameters, any non-constant signal is enough to estimate the delay.

- For systems with a small range of parameter variations, the Estimated Model VRE (EMVRE) is introduced. Both the system parameters and delay are estimated interactively. The delay estimation is unbiased even if the parameter
estimates are biased due to the existence of coloured noise.

- For systems of large or unknown parameter variations, the Fixed Model VRE (FMVRE) is introduced. The delay estimation is independent of the parameter estimation. The delay estimation is unbiased for first and second-order systems and for higher-order systems of pole excess equal to one. For other systems, the delay estimation converges to a value close to the Ziegler-Nichols approximation.

- The FMVRE delay estimate is bounded and unbiased in the presence of coloured noise.

- The excitation richness required for delay estimation alone is less than that required to estimate the parameters if the delay is known. Excitation richness as low as one sinusoid can be enough to estimate the delay correctly.

- The estimated model is stable if the system is stable even in the presence of model order mismatch.

- If fractional delay exists, the FMVRE may have a bias of one in the delay estimation. This bias can easily be corrected by extending the model numerator polynomial by one.

- Unlike the model parameters, the model delay cannot lose identifiability in closed-loop. The noise has no effect on the delay estimation in open-loop identification. In closed-loop identification, a delay estimation bias may result because of the noise. To avoid this bias, different options are given.

- VRE is linked with controller designs and the adaptive versions of pole placement, Dahlin and generalized predictive controllers successfully applied to systems with
variable parameters and delay. The VRE-based GPC provides the same closed-loop performance over the range of delay variation. This was not possible in the original GPC.

- Experiments for closed-loop and open-loop delay estimation are performed. Adaptive Dahlin/FMVRE is used to control the temperature in an experiment having long delay. The results show fast adaptation and good closed-loop performance over a wide range of delay variation. In a second experiment EMVRE is used to model a Reject Pulp Refiner. Although the measurements are very noisy, the delay estimation converges in less than 20 samples to the value obtained by offline analysis. The experiments demonstrate the applicability of the VRE delay estimation methods.

1.4 Thesis Outline

Chapter 2 shows the importance of correct delay estimation. The standard delay estimation methods are reviewed and the limitations of every method are highlighted.

Chapter 3 describes the idea of Variable Regression and how the EMVRE can be generated by modifying any standard parameter estimator. It is shown that under certain conditions, the delay estimation converges only to the correct value.

Chapter 4 introduces the FMVRE and shows how the model parameters are selected. Proofs of the estimation convergence, bounds, robustness and consistency in the presence of coloured noise are given. The excitation conditions required for accurate delay estimation are derived and it is shown that one sinusoid is sufficient excitation.
Chapter 5 presents the adaptive versions of pole placement, Dahlin and generalized predictive controllers designed based on the VRE. The analysis of delay estimation in closed-loop is also presented.

Chapter 6 shows the experimental results of applying the VRE to a laboratory temperature control experiment subject to a wide range of delay variation. Results from an identification experiment on an industrial reject pulp refiner are also presented.

Chapter 7 summarizes the results of the thesis, and gives suggestions for further research.
Chapter 2

Introduction to Delay Estimation

Many systems have a time-varying delay as well as time-varying dynamics. A good on-line estimation algorithm is required in these cases to track both the parameters and the delay changes. On the other hand, most control designs are sensitive to wrong delay assumption or estimation and can easily go unstable in such a case. This fact emphasizes the need for a fast and accurate on-line process delay estimation. One method to overcome the problem of unknown delay is to use an extended horizon controller or a generalized predictive controller designed for the maximum expected delay (e.g. Clarke et al. 1987). A major drawback of such a method is the unnecessarily detuned response that results.

Consider the continuous system that has the transfer function:

\[ G(s) = \frac{Y(s)}{U(s)} = K_{ss}e^{-T_d s} \prod_{i=1}^{n} \frac{(s/\omega_{zi} + 1)}{(s/\omega_{pi} + 1)} \]  

where:

- \( K_{ss} \) is the steady state gain,
- \( T_d \) is the delay,
- \( n \) is the number of poles,
- \( \omega_{pi} \) is the \( i \)'th pole corner frequency, and \( \omega_{zi} \) is the \( i \)'th zero corner frequency.

The transfer function \( G(z) \) is obtained using the \( Z \) transform and a zero order-hold
as:

$$G(z) = (1 - z^{-1}) \mathcal{Z} \left\{ \frac{G(s)}{s} \right\}$$  \tag{2.2}

and has the form:

$$\frac{Y(z)}{U(z)} = G(z) = z^{-k} \frac{b_0 z^{-1} + b_1 z^{-2} + \cdots + b_{n-1} z^{-n}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}}$$  \tag{2.3}

where $k = \text{integer}(T_d/T_s)$ is the delay in samples, and $T_s$ is the sampling time. Equation 2.3 can be written in the vector form:

$$y(t + 1) = \Phi(t) \Theta,$$  \tag{2.4}

$$\Phi(t) = [-y(t), \ldots, -y(t - n + 1), u(t - k), \ldots, u(t - k - n + 1)],$$  \tag{2.5}

$$\Theta = [a_1, \ldots, a_n, b_0, \ldots, b_{n-1}]^T$$  \tag{2.6}

$\Phi(t)$ is the regression vector and $\Theta$ is the parameter vector. In this study only the elements in the parameter vector are called the “parameters” of the system. The system delay does not show up as an element in the parameter vector, but it shows up in the structure of the regression vector. Our objective is to estimate $k$ with only prior knowledge about its bounds:

$$k \in [k_{\text{min}}, k_{\text{max}}]$$  \tag{2.7}

and $\Theta$ can be either known or unknown. The following example demonstrates the importance of a correct delay estimation for control.

**Example 2.1**

Consider the first-order system:

$$G(z) = z^{-k} \frac{b_i}{z - a_i}$$
Chapter 2. Introduction to Delay Estimation

which is estimated as:

\[ \hat{G}(z) = z^{-\hat{k}} \frac{\hat{b}}{z - \hat{\alpha}} \]

and the required closed-loop transfer function is:

\[ G_r(z) = z^{-\hat{k}} \frac{\beta}{z - \alpha} \]

The error feedback controller \( G_{fb}(z) \) required to obtain this closed-loop function is given by:

\[ G_{fb}(z) = \frac{1}{G(1 - G_r)} \]

The actual closed-loop transfer function \( G_{cl}(z) \) equals the required one \( G_r \) only if we have a correct model: \( \hat{\alpha} = \alpha, \hat{\beta} = \beta \) and \( \hat{k} = k \). The root-locus technique can be used to study the effect of having errors in estimating \( a \) and \( b \). It is well known that a small error in those parameters changes the location of the closed-loop poles but the poles can still be inside the unit circle. On the other hand, the root-locus technique cannot be used to study the closed-loop pole locations as a function of the delay estimate because the characteristic equation \( 1 + G_{fb}G = 0 \) is not linear in \( \hat{k} \) and because a change in \( \hat{k} \) can also change the order of the characteristic equation. There is only a limited number of possible \( \hat{k} \) defined by 2.7 and the roots of the characteristic equation can be calculated for all possible \( \hat{k} \).

Assuming perfect knowledge of the system parameters \( \hat{\alpha} = a, \hat{\beta} = b \), the closed-loop transfer function is then:

\[ G_{cl} = \frac{\beta}{z - \alpha + \beta(z^{\hat{k}} - z^{\hat{k}})} \]
Consider for example that $\alpha = .36, \beta = .64$ and $k_{\text{min}} = 1, k_{\text{max}} = 5$. The possible combinations of $k$ and $\hat{k}$ give:

\[
\begin{array}{ccccc}
\hat{k} & 1 & 2 & 3 & 4 & 5 \\
1 & - & - & * & * & * \\
2 & - & - & - & * & * \\
3 & - & - & - & - & * \\
4 & * & * & - & - & - \\
5 & * & * & * & - & - \\
\end{array}
\]

where "—" denotes a stable closed-loop system and "*" denotes an unstable or very oscillatory closed-loop system. Most of the stable combinations have highly oscillatory poles which give a poor closed-loop response. As the difference $|k - \hat{k}|$ increases, the system can be easily driven to instability, even that we know the parameters $a$ and $b$ perfectly. One way to assure a stable system is to ask for a slower closed-loop response. If the required closed-loop system is twice as slow as the case above, we get $\alpha = .6, \beta = .4$, and for these parameters we have a stable closed-loop for any combination of $k$ and $\hat{k}$. So, a detuned controller will be robust for errors in the estimated delay, but at the price of a slower closed-loop response. It is clear that a correct delay estimation is essential for a good closed-loop performance.

2.1 Methods of Delay Estimation

In this section, the important delay estimation methods are reviewed. In on-line estimation, we have the extended numerator polynomial method and the rational approximation method. In off-line estimation we have the cross-correlation method.
2.1.1 The Extended Numerator Polynomial Method

In this method, we replace the numerator polynomial in Equation 2.3 by an extended one and the expected output can be written in the form equivalent to that in Equation 2.4 as follows:

\[
y_e(t + 1) = \Phi_e(t)\Theta_e,
\]
\[
\Phi_e(t) = [-y(t), \ldots, -y(t - n + 1), u(t - k_{\min}), \ldots, u(t - k_{\max} - n + 1)],
\]
\[
\Theta_e = [a_1, \ldots, a_n, b_{k_{\min}}, \ldots, b_{k_{\max} + n - 1}]^T
\]

The suffix \( e \) stands for the extended polynomial. There are additional \((k_{\max} - k_{\min})\) elements in the parameter vector. The first \((k - k_{\min})\) elements and the last \((k_{\max} - k)\) elements in the parameter vector should be zeros in order to have \( y_e = y \).

In order to estimate \( k \) and \( \Theta \), we have to estimate the extended vector \( \Theta_e \) and conclude the value of \( k \) from the first zero elements. The method is heavily used in practice (e.g. Biswas and Singh, 1978; Vogel and Edar, 1980; Kurz and Goedecke, 1981; Gerry et al., 1983; Chien et al., 1984; Chandra et al., 1985; Clough and Park, 1985; Lammers and Verbruggen, 1985 and in many others). The method has the following drawbacks:

1. This method essentially requires an unbiased estimate of \( \Theta_e \). This means that the input signal should be rich enough to identify the \( 2n \) system parameters plus the additional \((k_{\max} - k_{\min})\) elements.

2. If there is coloured noise, then the estimated parameters will be biased and the delay cannot be concluded. This means that the noise model should also be estimated which means additional parameters to estimate and a slower convergence. Thus, the Recursive Least Squares (RLS) parameter estimator cannot
be used and the Approximate Maximum Likelihood (AML) estimator should be used instead.

3. The number of parameters in the extended numerator polynomial increases linearly with the system delay or the sampling rate. This means slower parameter convergence and longer time to identify the delay.

2.1.2 The Rational Approximation Method.

In this method the term $e^{-T_d s}$ in Equation 2.1 is approximated by a low-order rational approximation $G_d(s)$, that is:

$$
G(s) \approx K_{ss} G_d(s) \prod_{i=1}^{n} \frac{(s/\omega_{zi} + 1)}{\prod_{i=1}^{n} (s/\omega_{pi} + 1)}
$$

The approximations include Padé approximation (Robinson and Soudack, 1970 and Gabay and Merhav, 1976), Walsh functions (Rao and Siuakumar 1979 and Rao 1983), Tchebycheff approximations (Knowles and Emre 1985), all pole approximation (Gawthrop and Nihtila, 1985), Laguerre series (Zervos, 1988a,b and Zervos et al. 1990) and all-pass filters (De Souza et al. in 1988). The approximation is selected such that it provides a good approximation of the system in the bandwidth of interest. The discrete time model of the system will not have an explicit value of the correct delay because the original system parameters and the delay approximation parameters cannot be decoupled.

The rational approximation has the following drawbacks:

1. The structure of the rational function is not kept after sampling which means that the delay cannot be distinguished from the process dynamics.

2. Most of the approximations use non-minimum phase functions which restrict the control design choices to indirect methods that preserve the unstable zeros.
3. The poles of the delay approximation must be close to the system poles to insure that the sampling rate is suitable for identifying both the system poles and the delay approximation poles. This means that the order of the approximation should increase linearly with process delay.

2.1.3 The Cross-Correlation Method

The main idea here is that the impulse response of the system is the same as the input-output cross-correlation if the input is white. The delay is then obtained as the time of first non-zero output on the system impulse response. The method is used mainly in signal analysis (Knapp and Carter, 1976; Hassab and Boucher, 1979) or to estimate the bound on the delay variation $[k_{\text{min}}, k_{\text{max}}]$ (Allison et al, 1989). The disadvantages of this method are:

1. The input must be white to have an impulse auto-correlation function. Any other input does not have a zero auto-correlation at all non-zero time shifts. The resulting cross-correlation function is then not zero for time shift less than the delay. And consequently the estimated delay will always be less than the true one.

2. The limit between what is considered zero and what is considered non-zero for the cross-correlation function should be very carefully defined. This limit allows a non-white input to be used, but on the other hand this limit depends mainly on the system which is generally unknown.

3. The method is mainly for open-loop estimation because it does not force causality on the estimated system and the feedback path between the input and the output (the controller) may be estimated instead of the feedforward path (the system).
4. The method is an off-line method because a large number of measurements should be collected before the cross-correlation function can be generated and the delay can be estimated.

2.2 Conclusions

In this chapter, we demonstrated the importance of having an accurate delay estimation and showed that wrong delay estimation can easily destabilize the closed-loop. The standard methods of delay estimation are reviewed. The extended numerator polynomial, the rational approximation and the cross-correlation methods are discussed and their disadvantages are focused on.
In this chapter, we present a new method for the simultaneous recursive estimation of both process parameters and delay. The new algorithm is a modification of any standard recursive parameter estimation algorithm and thus is easily applied to any estimation algorithm such as Least Squares, Instrumental Variables, Maximum Likelihood, ... etc. and their extensions. The new algorithm requires minimum additional computations compared to the original algorithm. The delay is explicitly estimated and the parameter vector has the same length as for the known delay system and no additional excitation is required. This means clear separation between the system dynamics and the delay.

### 3.1 Variable Regression

The basic idea in variable regression delay estimation is to use a first-order model to estimate the delay. The estimated delay is then used to modify the structure of the regression vector and a variable regression is generated. Let the system be described by the ARMAX equation:

$$y(t) = \Phi(t, k)\Theta + w(t)$$  \hspace{1cm} (3.9)

where:

$$\Phi(t, k) = [-y(t-1), ..., -y(t-n), u(t-k-1), ..., u(t-k-n)]$$  \hspace{1cm} (3.10)

$$\Theta = [a_1, ..., a_n, b_0, ..., b_{n-1}]^T$$  \hspace{1cm} (3.11)
Chapter 3. Delay Estimation Using Variable Regression

$k$: system time delay

$n$: number of poles

$w(t)$: coloured noise.

The parameter vector $(\Theta)$ is a function of the parameters, while the measurement vector $(\Phi)$ is a function of the time delay. Since many industrial processes are overdamped and can well be approximated as first-order plus delay, the model used is first-order plus delay. It will be shown that the first-order model is enough to estimate the delay even if the system dynamics are represented by a higher-order model. The estimated delay can then be used in the estimator of the high-order model parameters.

Let the model parameters and the estimated time delay be $\Theta_m$ and $\hat{k}$ respectively. The predicted output is given by:

$$\hat{y} = \hat{\Phi}(t, \hat{k}) \Theta_m$$

(3.12)

where:

$$\Theta_m = [a_m, b_m]^T$$

(3.13)

$$\hat{\Phi}(t, \hat{k}) = [-y(t-1), u(t-\hat{k}-1)]$$

(3.14)

where $0 \leq -a_m \leq 1$. $\hat{k}$ can take any integer value in the horizon $[k_{\text{min}}, k_{\text{max}}]$. The structure of the regression vector $\hat{\Phi}$ is variable depending on the estimated value $\hat{k}$. This variation of the regression vector characterizes this method and hence the designation is "Variable Regression Estimator (VRE)". The prediction error is defined as:

$$\epsilon(t, \Theta_m, \hat{k}) = y(t) - \hat{y}(t) = y(t) - \hat{\Phi}(t, \hat{k}) \Theta_m$$

(3.15)

The performance index is chosen to be:

$$J(t, \Theta_m, \hat{k}) = \frac{1}{t} \sum_{i=0}^{t} [\epsilon(i, \Theta_m, \hat{k})]^2$$

(3.16)
Chapter 3. Delay Estimation Using Variable Regression

The performance index is a function of both the model parameters and delay. So in order to minimize the performance index, Equation 3.16 should be minimized with respect to both the model parameters and time delay. The solution will identify $\Theta_m$ and $\hat{k}$. This means that:

$$\frac{\partial J(\Theta_m, \hat{k})}{\partial \Theta_m} = 0$$  \hspace{1cm} (3.17)

$$J(\Theta_m, \hat{k}) = \min[J(\Theta_m, k_i)] \forall k_i \in [k_{\min}, k_{\max}]$$  \hspace{1cm} (3.18)

Equation 3.18 is not written in the partial derivative form because $\hat{k}$ takes only discrete values. This characteristic of the delay equation makes the solution of 3.18 robust to modelling error in $\Theta_m$ and/or the existence of noise (white or coloured). If the model has the correct system parameters and in the absence of noise, Equation 3.18 will have a minimum of zero. If there is a small modelling error and/or process noise, Equation 3.18 alone still can be enough to identify the delay correctly. The following Theorem shows that in the case of known parameter system, the delay can be accurately estimated using Equation 3.18 alone.

**Theorem 3.1 (Known System parameters)** *For any stable system of known parameters ($\Theta_m = \Theta$), if the measurement noise is uncorrelated with the system input, then $J(\hat{k})$ is minimum when $\hat{k} = k$.*

Proof: Let:

$$y(t) = \Phi(t, k)\Theta + w(t)$$

$$\hat{y}(t) = \hat{\Phi}(t, \hat{k})\Theta$$

then:

$$e(t, \hat{k}) = (\Phi(t, k) - \hat{\Phi}(t, \hat{k}))\Theta + w(t)$$
The performance index is then:

\[ J(t, \hat{k}) = \frac{1}{t} \sum_{i=0}^{t} [\varepsilon(i)]^2 \]

\[ = \frac{1}{t} \sum_{i=0}^{t} [w^2(i)] + [(\Phi(i, k) - \hat{\Phi}(i, \hat{k}))\Theta]^2 + [2w(i)(\Phi(i, k) - \hat{\Phi}(i, \hat{k}))\Theta] \]

As the number of measurements increases, the different terms in the last equation can be replaced by their expected values, then:

\[ J(\infty, \hat{k}) = \lim_{t \to \infty} J(t, \hat{k}) \]

\[ = E(w^2(i)) + E((\Phi(i, k) - \hat{\Phi}(i, \hat{k}))\Theta)^2 + 2E(w(i)(\Phi(i, k) - \hat{\Phi}(i, \hat{k}))\Theta) \]

The term \( E(w^2(i)) \) is by definition equal to \( r_w(0) \), the auto-correlation function of \( w \) at zero shift. The last term should be zero because it is a summation of terms of the form \( E(w(i)u(i-j)) \) and this is by definition equal to \( r_{uw}(j) \), the cross-correlation function of \( w \) and \( u \) at shift \( j \) and this cross-correlation is zero at all shifts because the two signals are uncorrelated. The second term is the squared summation of the input auto-correlation function values at different shifts and this squared summation cannot be negative.

From the above analysis, we conclude that \( J(\hat{k}) = r_w(0) \) for \( \hat{k} = k \), and \( J(\hat{k}) \geq r_w(0) \) for \( \hat{k} \neq k \). There are only two cases for which \( J(\hat{k}) = r_w(0) \) and \( \hat{k} \neq k \). The first case is when the input has a flat auto-correlation function \( (r_u(\tau) = r_u(0)\forall\tau) \). This corresponds to a constant level input \( (u(t) = constant, \Phi(t, k) = \Phi(t, \hat{k})) \). The system will have a steady output and will look the same for any value of the delay. Any input change is sufficient for correct delay estimation. However, a richer input will have a sharper auto-correlation function and a sharper minimum of \( J(\hat{k}) \) at \( \hat{k} = k \). The second case for \( J(\hat{k}) = r_w(0) \) and \( \hat{k} \neq k \) is when the input signal is periodic. In this case \( J(\hat{k}) \) is also periodic and have the same period \( T_e \), i.e. \( J(\hat{k} + T_e/T_e) = J(\hat{k}) \). The input signal
should be so slow that $J(\hat{k})$ has only one minimum in the possible range $[k_{\text{min}}, k_{\text{max}}]$. $T_e$ must then satisfy that $T_e/T_s > (k_{\text{max}} - k_{\text{min}})$. This condition defines the lower limit of the excitation period. This condition is very general and not restricted to systems of known parameters. □

The result of this theorem is valid for any system order, provided that the model has the same order and parameters as the system. Any non-constant input is sufficient and for periodic input the condition given above should be satisfied. The case of unknown parameters is studied in the following theorem for the first-order systems.

**Theorem 3.2 (Unknown System Parameters)** If the input is white and uncorrelated with the coloured measurement noise, then minimizing $J$ gives an unbiased delay estimation for a first-order stable system of unknown parameters.

Proof: Let:

$$y(t) = ay(t-1) + bu(t-1 - k) + w(t)$$

then,

$$\hat{y}(t) = a_m y(t-1) + b_m u(t-1 - \hat{k})$$

then,

$$\epsilon(t, \hat{k}) = (a_m - a) y(t-1) + bu(t-1 - k) - b_m u(t-1 - \hat{k}) + w(t)$$

A similar analysis to that in the known parameters case gives:

$$J(\infty, \Theta_m, \hat{k}) = [(a_m - a)^2 r_y(0) + (b^2 + b_m^2)r_u(0) + r_w(0) + 2(a_m - a)(br_{uy}(k) + r_{wy}(1))]$$

$$-2 b_m [(a_m - a)r_{uy}(\hat{k}) + br_u(\hat{k} - k)]$$

The terms in the first square brace do not depend on $\hat{k}$, and so, only the two terms in the second square brace should be considered in the minimization. For a white
excitation:

\[
\begin{align*}
    r_u(\tau) &= \begin{cases} 
        r_u(0) & \text{if } \tau = 0 \\
        0 & \text{if } \tau > 0
    \end{cases}, \\
    r_{uv}(\tau) &= \begin{cases} 
        0 & \text{if } \tau \leq k \\
        br_u(0)|a|^{r+1+k} & \text{if } \tau > k
    \end{cases}
\end{align*}
\]

and since \(|a_m - a| < 1\), then \((a_m - a)r_{uv}(\hat{k}) < br_u(0)\) and \(J(\hat{k})\) is minimum at \(\hat{k} = k\), and this proves the theorem. \(\Box\)

It is to be noted that having a white input is not a necessary condition, but rather a sufficient condition. As the difference \(|a_m - a|\) decreases, the minimum of \(J(\hat{k})\) occurs at \(\hat{k} = k\) for other inputs not as rich as the white input. It must be noted that we assume a positive value for \(a\), which is the case for any physical system.

Example 3.1

Consider the first-order system:

\[
y(t) = ay(t-1) + q^{-k}bu(t-1) + w(t)
\]

where \(a = -0.5\), \(b = 0.5\) and \(k = 3\) and \(u\) is a step input. Three tests are carried out and the results are summarized in Figure 3.1. The three cases studied are:

1. Correct parameter estimates \((\Theta_m = \Theta)\) and no noise \((w = 0)\). Equation 3.18 has a sharp minimum of zero. Figure 3.1 shows a family of curves, each of them represents a specific time instant with data collected only up to that point. Curves representing time intervals less than the delay will have a flat minimum and reflect the poverty of the information collected.

2. 50% error in the estimated parameters \((\hat{a} = -0.25, \hat{b} = 0.75)\) and no noise. Equation 3.18 alone is still able to identify the delay accurately, although the input is not white. The minimum is not zero any more and this usually indicates the inaccuracy in the estimated parameters.
Figure 3.1: Minimization of $J$
3. As in the case above but with additional measurement noise \( w(t) = N(0, .1) \).

Even under this condition, Equation 3.18 alone is still enough to identify the delay accurately.

The last two theorems imply that there is a strong correlation between the excitation requirement of the input signal and the uncertainties of the parameters. The more accurate the parameter estimates are, the less is the excitation required for correct delay estimation. Since the system parameters are generally unknown, we have either one of two choices:

1. Estimating the model parameters that best describe the system. In this case Equations 3.17 and 3.18 should be solved simultaneously. This case will be called the **Estimated Model Variable Regression Estimation (EMVRE)**. Both the model parameters and the delay are to be estimated recursively.

2. Using fixed model parameters to estimate the delay by Equations 3.18 only. The fixed model should be chosen such that it is most likely to have correct delay estimation. This method will be called the **Fixed Model Variable Regression Estimation (FMVRE)**. The delay is estimated recursively and the estimated value is fed to the variable regression vector of another model used to describe the system and whose parameters are to be estimated recursively.

### 3.2 The Estimated Model Variable Regression Estimation (EMVRE)

In this case both the model parameters and the delay are estimated on-line. Since good estimation of the model parameters results in a good delay estimation, it is logical to use the most recent parameter estimates to estimate the delay. Similarly, since good parameter estimation requires good delay estimation, it is logical to use the most
recent delay estimate to estimate the parameters. The implementation of the model parameters and the delay estimation is carried out in two steps. In the first step the standard Recursive Least-Squares estimator is used to estimate the parameters using the last estimated delay. The next step is to estimate the delay using the last estimated parameters. The two steps are to be carried out once every sample. The steps of the modified version of the standard Recursive Least-Squares estimator will be:

1. Start with initial guess: \( \Theta_m = \Theta_0 \) and \( k_{\text{min}} \leq \hat{k} \leq k_{\text{max}} \).

2. Generate the corresponding regression vector \( \hat{\Phi}(\hat{k}) \).

3. Update the parameter estimation \( (\Theta_m) \) using RLS.

4. Update the performance index values \( (J) \) over the horizon \( [k_{\text{min}}, k_{\text{max}}] \) using the recent parameter estimates.

5. Estimate the delay \( \hat{k} \) that minimizes the performance index.

6. New sample: go to 2 above.

The equations required to implement those steps are as follows:

\[
K(t + 1) = P(t)\hat{\Phi}^T(t + 1, \hat{k}(t))/[\lambda + \hat{\Phi}(t + 1, \hat{k}(t))P(t)\hat{\Phi}^T(t + 1, \hat{k}(t))] \tag{3.19}
\]

\[
\Theta_m(t + 1) = \Theta_m(t) + K(t + 1)(y(t + 1) - \hat{\Phi}(t + 1, \hat{k}(t))\Theta_m(t)) \tag{3.20}
\]

\[
P(t + 1) = [I - K(t + 1)\hat{\Phi}(t + 1, \hat{k}(t))]P(t)/\lambda \tag{3.21}
\]

\[
J(t + 1, k_i) = \lambda J(t, k_i) + [y(t + 1) - \hat{\Phi}(t + 1, k_i)\Theta_m(t)]^2 \forall k_i \in [k_{\text{min}}, k_{\text{max}}] \tag{3.22}
\]

\[
J(t + 1, \hat{k}(t + 1)) = \min_{k_i \in [k_{\text{min}}, k_{\text{max}}]} J(t + 1, k_i) \tag{3.23}
\]

where \( P \) is the parameter covariance matrix, \( K \) is the gain matrix, and \( \lambda \) is the forgetting factor.
Equations 3.19-3.21 are the standard Least-Squares estimator equations for a constant \( \hat{k} \). The additional two equations are the delay estimation equations. The implementation of the additional equations requires minimum storage and computation time as they contain only simple multiplications and additions and a simple search routine. Equations 3.22 and 3.23 can be simply added to any recursive estimator.

Computer simulations show that the delay estimation does not have to change with every change in the parameter estimation. In other words, the delay estimation will not change at every sample as the parameters will. As more data are collected, the parameter changes decrease towards the final convergence values and at the same time the period at which the delay estimation is not changing increases. As the parameters converge to constant values, the delay estimation will reach its correct value and leave the parameter estimation to converge to the correct values uninterruptedly. This characteristic suggests resetting the parameter covariance matrix (\( P \)) every time the delay estimation changes to speed up the parameter estimation convergence.

If the delay estimation converges, the parameter estimation will have the same steady-state properties as the original estimator. It is difficult to derive the general convergence properties of this method because of two reasons. The first is the mutual interaction between the delay and the parameter estimation which means that the delay and parameter estimation equations should be solved simultaneously. The second reason is that the delay updating equation is not in a difference equation form as in the case of the parameter estimation equations. However, for a first-order system we can prove that if the algorithm converges, it will converge only to the correct delay. In other words there is only one convergence point for the algorithm.

**Theorem 3.3 (Convergence of the EMVRE)** For a first-order stable system excited by a white input, if the EMVRE delay estimation converges, it converges only to
Proof: Let the algorithm converge to the values $a_m, b_m, \hat{k}$, then from Theorem 3.2 we have $\hat{k} = k$ provided that the excitation signal is rich enough. □

The estimated parameters are essentially those obtained by RLS using the correct delay. These parameter estimates can be biased in the presence of coloured noise. However the Approximate Maximum Likelihood can be modified with the variable regression technique to guarantee unbiased parameter estimates.

Although the result of this theorem is restricted to white input, computer simulation and real-time experiments show that correct delay estimation can be obtained with a much smaller level of excitation. The following examples show the characteristics of the EMVRE algorithm. In the first example, a first-order deterministic model is considered. The parameters and the delay have large simultaneous changes. The second example shows the EMVRE working in an adaptive pole placement loop. The third example shows the estimator behaviour over a wide range of sampling rate variations. In the last example, the EMVRE is used as a method of model reduction.

**Example 3.2 (Variable Parameters and Delay)**

Consider the system:

$$y(t) = -ay(t-1) + q^{-k}bu(t-1)$$

The input $u(t)$ is a square wave with period of 40 sample. $a, b, k$ change with time as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>a</th>
<th>b</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 &lt; t &lt; 200$</td>
<td>-.50</td>
<td>.50</td>
<td>3</td>
</tr>
<tr>
<td>$200 &lt; t &lt; 400$</td>
<td>-.25</td>
<td>.75</td>
<td>1</td>
</tr>
<tr>
<td>$400 &lt; t &lt; 600$</td>
<td>-.75</td>
<td>.25</td>
<td>5</td>
</tr>
</tbody>
</table>
Chapter 3. Delay Estimation Using Variable Regression

The simultaneous large changes in the parameters and delay represent a good evaluation test for the proposed estimation method. The results in Figure 3.2 show a fast and accurate tracking of the changing parameters and delay.

Example 3.3 (Adaptive pole placement)

Many adaptive controllers have been simulated successfully using the EMVRE. Some adaptive controllers are very sensitive to correct delay assumption or estimation. The pole-placement controller is one of those delay-sensitive controllers. The system used in Example 3.2 is now reused with an adaptive pole placement controller with a closed-loop time constant of two samples. The controller design can be found in Chapter 5 where detailed analysis of the adaptive controller design is given. The simultaneous large changes in the parameters and delay would drive the system to instability if a non-adaptive controller is used. The results of the adaptive controller based on EMVRE are shown in Figure 3.3.

Example 3.4 (Variable Sampling Time)

Consider the system:

\[ G(s) = e^{-T_d s} \frac{1}{\tau s + 1} \]

Where \( T_d = 1 \text{sec.}, \tau = 1 \text{sec.} \). The sampled system has the form of Example 3.2. The sampling rate directly affects the values of the parameters and the delay. Three sampling rates are used to test the estimator. The model equation has the following parameters as the sampling rate \( (F_s) \) changes:

<table>
<thead>
<tr>
<th>( F_s )</th>
<th>a</th>
<th>b</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Hz</td>
<td>-.3678</td>
<td>.6321</td>
<td>1</td>
</tr>
<tr>
<td>5 Hz</td>
<td>-.8187</td>
<td>.1813</td>
<td>5</td>
</tr>
<tr>
<td>10 Hz</td>
<td>-.9048</td>
<td>.0951</td>
<td>10</td>
</tr>
</tbody>
</table>
Figure 3.2: EMVRE of a first order system.
Figure 3.3: Estimation and adaptive P-P control using the EMVRE.
The results in Figure 3.4 show excellent estimation of the true values of the parameters and of the time delay for the different sampling rates. The input signal used was a square wave.

**Example 3.5 (High order System)**

The system has a fourth-order transfer function:

\[ G(s) = \frac{W_1^2}{s^2 + 2\eta_1 W_1 s + W_1^2} \frac{W_2^2}{s^2 + 2\eta_2 W_2 s + W_2^2} \]

Where: \(\eta_1 = 0.5, W_1 = 1.0\text{rad/sec}, \eta_2 = 0.75, W_2 = 2.0\text{rad/sec}\). The step response of the fourth-order model and the second order dominant model are shown in Figure 3.5 as \(Y_4\) and \(Y_2\) respectively. It is clear that the fourth-order model response can be represented by a delayed second order one. The sampling rate is 10Hz and a second order system with delay is estimated using a square wave input. The step response of the estimated model is shown in Figure 3.5 as \(\hat{Y}\). This simulation shows that the dominant poles are not enough to describe the system and a delay is a good approximation of the unmodelled dynamics. This process is the opposite of approximating the delay by a rational function and is a common industrial practice.

### 3.2.1 Advantages and disadvantages of the EMVRE

The EMVRE has unique advantages over both the extended numerator polynomial and the rational approximation. The main advantages of the EMVRE are:

1. Since no parameters are added to the process dynamics model, the parameter convergence is fast and no additional richness conditions are required from the input signal to estimate the model parameters.

2. Unbiased delay estimation can be obtained in presence of coloured noise even with biased model parameters.
Figure 3.4: EMVRE at different sampling rates.
Figure 3.5: EMVRE of a high order system.
3. The additional equations required to estimate the delay are easy to implement with minimum code, computation time and storage.

The EMVRE main disadvantages are:

1. There is no proof that the algorithm will always converge.

2. There is no general result on the method behaviour with system order higher than one.

3. The exact excitation requirement for correct delay estimation is not known.

The disadvantages of the EMVRE could be overcome with the introduction of the FMVRE in the next chapter.

3.3 Conclusions

In this chapter, the new idea of variable regression is presented and used to develop the Estimated Model VRE. The VRE is combined with Least-Squares parameter estimator to form recursive estimation of both the system parameters and delay. For systems with known parameters, the delay is estimated with the minimum possible excitation. For first-order systems, the EMVRE can converge only to the true values when rich excitation is used.
Chapter 4

The Fixed Model Variable Regression Estimation (FMVRE)

The FMVRE uses a fixed parameter model to estimate the delay. It will be shown that the best results are obtained when the first-order model pole lies on the unit circle (an integrator). The method in this special case is shown to be equivalent to maximizing the cross-correlation between the input and the output increments. This is compared with the standard cross-correlation method and the advantages are clarified.

Consider the stable system:

\[
\frac{Y(z)}{U(z)} = G(z) = z^{-k} \frac{b_0 z^{-1} + b_1 z^{-2} + \cdots + b_{n-1} z^{-n}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}}
\] (4.24)

where: \( k \geq 0 \) is the system time delay, and \( n \) is the system order.

\( G(z) \) can be expanded as follows:

\[
G(z) = z^{-k} \left[ \frac{c_1}{z-p_1} + \frac{c_2}{z-p_2} + \cdots + \frac{c_n}{z-p_n} \right]
\] (4.25)

where \( c_i \) and \( p_i \) are real or complex numbers and \( |p_i| \leq 1 \ \forall \ i \in [1, n] \).

\( G(z) \) can be expanded as an infinite series:

\[
G(z) = z^{-k-1} \left[ c_1 \left( 1 + p_1 z^{-1} + p_1^2 z^{-2} + p_1^3 z^{-3} + \cdots \right) \\
+ c_2 \left( 1 + p_2 z^{-1} + p_2^2 z^{-2} + p_2^3 z^{-3} + \cdots \right) \\
\vdots \\
+ c_n \left( 1 + p_n z^{-1} + p_n^2 z^{-2} + p_n^3 z^{-3} + \cdots \right) \right]
\]
The impulse response coefficients of this system are:

$$y_{\text{impulse}}(t) = \begin{cases} 
0 & \text{if } (t - 1) < k \\
\sum_{i=1}^{n} c_i & \text{if } (t - 1) = k \\
\sum_{i=1}^{n} c_i p_i^{(t-1)-k} & \text{if } (t - 1) > k 
\end{cases}$$  \hspace{1cm} (4.26)

Using a first-order plus delay model, the estimated output is defined as:

$$\hat{y}(t) = q^{-1} a_m y(t) + q^{-k-1} b_m u(t)$$  \hspace{1cm} (4.27)

and $0 \leq a_m \leq 1$. The prediction error is given by:

$$\varepsilon^2 = (y(t) - \hat{y}(t))^2$$  \hspace{1cm} (4.28)

$$= (y(t) - a_m y(t-1) - b_m u(t-1-k))^2$$  \hspace{1cm} (4.29)

$$= (y^2(t) + a_m^2 y^2(t-1) + b_m^2 u^2(t-1-k) - 2a_m y(t)y(t-1) - 2b_m (y(t)u(t-1-k) - a_m y(t-1)u(t-1-k))$$  \hspace{1cm} (4.30)

The performance index $J(\Theta_m, \hat{k})$ is equivalent to the expected value of $\varepsilon^2$ and is given by:

$$J(\Theta_m, \hat{k}) = E(\varepsilon^2) = E_0 - 2b_m E_1$$  \hspace{1cm} (4.31)

where:

$$E_0 = (1 + a_m^2) r_y(0) + b_m^2 r_u(0) - 2a_m r_y(1)$$  \hspace{1cm} (4.32)

$$E_1 = r_{uy}(1 + \hat{k}) - a_m r_{uy}(\hat{k})$$  \hspace{1cm} (4.33)

and $r_y(0) = E(y^2(t)), r_{uy}(\hat{k}) = E(y(t)u(t - \hat{k}))$. The value of $E_0$ does not depend on $\hat{k}$, so in order to minimize $J(\Theta_m, \hat{k})$ we have to maximize $E_1$ with respect to $\hat{k}$. In the analysis given here we assume, without loss of generality, a positive static gain system.
For systems having negative static gains, $E_1$ should be minimized. From the impulse response coefficients of the system, we conclude that for white excitation:

$$
r_{uv}(1 + \hat{k}) = r_u(0) \begin{cases} 
0 & \text{if } \hat{k} < k \\
\sum_{i=1}^{n} c_i & \text{if } \hat{k} = k \\
\sum_{i=1}^{n} c_i p_i^{k-k} & \text{if } \hat{k} > k 
\end{cases}
$$

and

$$
r_{uv}(\hat{k}) = r_u(0) \begin{cases} 
0 & \text{if } \hat{k} < k \\
0 & \text{if } \hat{k} = k \\
\sum_{i=1}^{n} c_i p_i^{k-k-1} & \text{if } \hat{k} > k 
\end{cases}
$$

then:

$$
E_1 = r_u(0) \begin{cases} 
0 & \text{if } \hat{k} < k \\
\sum_{i=1}^{n} c_i & \text{if } \hat{k} = k \\
\sum_{i=1}^{n} c_i p_i^{k-k-1} - a_m c_i p_i^{k-k-1} & \text{if } \hat{k} > k 
\end{cases}
$$

The terms in the last equation are very similar to the terms of the impulse response coefficients of the system. In fact they are identical to the impulse response coefficients of the function:

$$
\frac{Y_1(z)}{U(z)} = G_1(z) = (1 - a_m z^{-1})G(z)
$$

which has the impulse response coefficients:

$$
y_{1 \text{ impulse}}(t) = \begin{cases} 
0 & \text{if } (t - 1) < k \\
\sum_{i=1}^{n} c_i & \text{if } (t - 1) = k \\
\sum_{i=1}^{n} c_i p_i^{(t-1)-k} - a_m c_i p_i^{(t-1)-k-1} & \text{if } (t - 1) > k 
\end{cases}
$$
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE)

The impulse response of the function $G_1(z)$ is identical to the response of the original system $G(z)$ when the input is:

$$U_1(z) = 1 - a_m z^{-1}$$  \hspace{1cm} (4.38)

$$u_1(t) = \begin{cases} 
1 & \text{if } 0 \leq t < 1 \\
-a_m & \text{if } 1 \leq t < 2 \\
0 & \text{if } 2 \leq t 
\end{cases}$$  \hspace{1cm} (4.39)

The last conclusion is important because it means that the value of $\hat{k}$ that maximizes $E_1$ can be determined from the system response to the input sequence defined in the last equation. This is illustrated in Figure 4.6 in which a general overdamped system having $k = 4$ is considered. The response of the system is shown and related to the function $E_1(\hat{k})$. In Figure 4.6 the maximum value of $E_1$ occurs at $\hat{k} \neq k$. That is because the estimated value of the delay depends on the values of $a_m$ and the sampling time.

From Equation 4.35, there are two conditions for unbiased delay estimation (i.e. $E_1$ is maximum at $\hat{k} = k$):

**Condition 4.1**

$$\sum_{i=1}^{n} c_i > 0$$

This condition means that the first non-zero value of the system impulse response is of the same sign as the steady state gain. This means that regardless of whether the system is minimum or non-minimum phase, the sampling rate is slow enough to have the first non-zero sample positive.

**Condition 4.2**

$$\sum_{i=1}^{n} c_i > \sum_{i=1}^{n} c_i p_i^j - a_m c_i p_i^{j-1}, \forall n \geq j > 0$$
Figure 4.6: Impulse response of $G_1$. 
This condition means that the first nonzero value of the system response to the input given by Equation 4.39 is the maximum of all the samples. The response will be influenced by the value of \( a_m \) that determines the amplitude of the negative input in Equation 4.39. The case when \( a_m = 0 \) is the most difficult one and is equivalent to maximizing the cross-correlation between the system input and output. In this case the condition for unbiased delay estimate is that the sampling rate is slow enough to have the first nonzero sample of the impulse response as the maximum value of all the samples. This is always true for first-order systems.

4.1 The Fixed Model Parameters

The other extreme case is when \( a_m = 1 \). In this case the input in the second sampling period is at its maximum negative value. This means that the second non-zero sample will have an amplitude much less than in the case \( a_m = 0 \) and consequently it is most likely that the first non-zero sample will be the maximum. The method is equivalent to maximizing the cross-correlation function between the system input and the output increments. The value of \( b_m \) is no longer important because we are concerned only with the maximization of \( E_1 \) which now takes the form:

\[
E_1 = r_{uy}(1 + \hat{k}) - r_{uy}(\hat{k}) = E((y(t) - y(t - 1))u(t - 1 - \hat{k})) = E(\Delta y(t)u(t - 1 - \hat{k})) = r_{u\Delta y}(1 + \hat{k})
\]

where:

\[
\Delta = (1 - q^{-1})
\]

Step disturbances and/or changes in the set point are automatically eliminated in \( \Delta y \) which means that they do not affect the delay estimation. From Equations 4.26 and
4.37, $E_1$ can be expressed as:

$$E_1(\hat{k} + 1) = y_{\text{impulse}}(t)$$
$$= \Delta y_{\text{impulse}}(t)$$
$$= \Delta \Delta y_{\text{step}}(t) = \Delta^2 y_{\text{step}}(t)$$
$$= (1 - 2q^{-1} + q^{-2})y_{\text{step}}(t)$$

(4.41)

The advantage of relating the function $E_1$ to the step response is that the step response does not change with changing the sampling time, while the impulse response does change. This leads us to the bounds of the delay estimation:

**Theorem 4.1** The time delay estimated by the FMVRE is bounded for any sampling rate.

Proof: For very large sampling time, the sampled step response is given by:

$$y_{\text{step}}(0) = 0$$

$$y_{\text{step}}(t > 0) = y_{ss} = \frac{b_0 + b_1 + \cdots + b_{n-1}}{1 + a_1 + \cdots + a_n}$$

Then:

$$E_1 = \begin{cases} 
0 & \text{if } \hat{k} + 1 = 0 \\
y_{ss} & \text{if } \hat{k} + 1 = 1 \\
-y_{ss} & \text{if } \hat{k} + 1 = 2 \\
0 & \text{if } \hat{k} + 1 > 2 
\end{cases}$$

Then the maximum of $E_1$ occurs when $\hat{k} = 0$, and this defines the lower bound of the estimated time delay. On the other hand, for very small sampling time, the operator $\Delta$ approximates the time derivative. The operator $\Delta^2$ approximates the second time derivative. This means that $E_1$ in Equation 4.41 is equivalent to the curvature of the step response curve:

$$E_1 \approx T_s^2 \frac{d^2}{dt^2} y_{\text{step}}(t)$$
Then the maximum of \( E_1 \) will occur at the point of maximum curvature of the step response and this defines the upper bound of the estimated delay. For any value of \( T_s \) such that: \( 0 < T_s < \infty \), the operator \( \Delta \) is finite and so is the value of \( E_1 \), and this proves the theorem. □

The upper bound on the delay estimate is closely related to the delay approximation given by the Ziegler-Nichols method (Åström 1988). In the Ziegler-Nichols approximation, the time delay is defined as the intercept of the maximum slope line with the time axis. For minimum phase systems, this time is usually very close to the time of maximum curvature as shown in Figure 4.7. The importance of this relation is that controllers tuned based on the Ziegler-Nichols approximation can now be tuned on-line to form an adaptive closed-loop.

### 4.2 Estimation of the System Parameters

The delay estimated using the FMVRE is independent of the other parameters of the system which are generally unknown. To estimate the parameters of the system, we need another model whose parameters are in \( \hat{\Theta} \). Standard RLS can be used to estimate \( \hat{\Theta} \) with a variable regressor \( \hat{\Phi}(\hat{k}) \). The estimation of the delay does not depend in any way on the vector \( \hat{\Theta} \), and the resulting decoupling between the delay and the parameter estimation gives the FMVRE several advantages over the EMVRE. The modification of the RLS is given here and similar modification can be applied to other estimators.

1. Start with initial guess: \( \hat{\Theta} = \Theta_0 \) and \( k_{\min} \leq \hat{k} \leq k_{\max} \).
2. Generate the corresponding regression vector \( \hat{\Phi}(\hat{k}) \).
3. Update the parameter estimation (\( \hat{\Theta} \)) using RLS.
4. Update \( E_1 \) over the horizon \([k_{\min}, k_{\max}]\).
Figure 4.7: Delay estimation using Ziegler-Nichols and FMVRE.
5. Look for $\hat{k}$ corresponding to the maximum of $E_1$.

6. New sample: go to 2

The equations required to implement the RLS/FMVRE are:

$$K(t+1) = P(t)\hat{\Phi}^T(t+1,\hat{k}(t))/[\lambda + \hat{\Phi}(t+1,\hat{k}(t))P(t)\hat{\Phi}^T(t+1,\hat{k}(t))]$$  \hspace{1cm} (4.42)

$$\hat{\Theta}(t+1) = \hat{\Theta}(t) + K(t+1)(y(t+1) - \hat{\Phi}(t+1,\hat{k}(t))\hat{\Theta}(t))$$  \hspace{1cm} (4.43)

$$P(t+1) = [I - K(t+1)\hat{\Phi}(t+1,\hat{k}(t))]P(t)/\lambda$$  \hspace{1cm} (4.44)

$$E_1(t+1, k_i) = \lambda E_1(t, k_i) + u(t - k_i)[y(t+1) - y(t)] \forall k_i \in [k_{\min}, k_{\max}]$$  \hspace{1cm} (4.45)

$$E_1(t+1, \hat{k}(t+1)) = \max E_1(t+1, k_i) \forall k_i \in [k_{\min}, k_{\max}]$$  \hspace{1cm} (4.46)

where $P$ is the parameter covariance matrix, $K$ is the gain matrix, and $\lambda$ is the forgetting factor.

The properties of the modified estimator are essentially the same as those of the standard estimator. The reason is that the delay estimation converges independently of the parameter estimation. As the delay estimation converges to its final value, the regressor is no longer variable and the estimator returns to its standard version. Important special cases of delay estimation result from the above general discussions.

### 4.3 First-Order Systems

Theorem 3.2 proves that for white excitation, the correct delay estimation is guaranteed. This result is of practical importance because many industrial processes behave as first-order systems with delay and for these systems unbiased delay estimation can be guaranteed. As stated earlier, having a white input is not a necessary condition but rather a sufficient condition. It will be proven later that for the FMVRE, one sinusoid can is a sufficient excitation for delay estimation. The following example demonstrates the FMVRE method with a first-order system of variable parameters and delay.
Example 4.1

Consider the system:

\[ G(s) = e^{-T_d s} \frac{1}{\tau s + 1} \]

The system time delay and time constant (in seconds) change with time as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>( T_d )</th>
<th>( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 &lt; t &lt; 100 )</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>( 100 &lt; t &lt; 200 )</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>( 200 &lt; t &lt; 300 )</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

The input \( u(t) \) is a square wave with period of 20 samples and \( T_s = 1 \text{sec} \). Figure 4.8 shows the estimation results in open-loop and in closed-loop using a pole placement controller with a closed-loop time constant of 1 sec. The controller design is given by Equation 2.8. The delay estimation converges to the true values in less than 20 samples.

4.4 Second-Order Systems

It is well known (Åström 1984-a) that for good parameter estimation and closed-loop control, the sampling time \( (T_s) \) is related to the rise time \( (T_r) \) by:

\[ T_r/2 > T_s > T_r/4 \]  \hspace{1cm} (4.47)

In this section we show that the FMVRE does not restrict the sampling rate choice for second-order systems with reasonable damping. Consider the general second-order system:

\[ G(s) = e^{-T_d s} \frac{bs + 1}{s^2/w_o^2 + 2\zeta s/w_o + 1} \]
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE) 44

Figure 4.8: FMVRE of a first order system
where $\zeta$ is the damping ratio and for simplicity assume that the system time delay is an integer multiple of the sampling time. The minimum sampling time required for unbiased delay estimation is shown in Figure 4.9 as a function of $\zeta$ for different values of $b$ and $w_0 = 1$. Given $b$ and $\zeta$, the minimum sampling time curve was obtained numerically by studying the system step response at different sampling rates and looking for the minimum sampling time that satisfies Condition 4.2. It is noted that for $b > 0$, the minimum sampling time required becomes very small and approaches zero as $b$ becomes > .1, as shown for the cases $b = 0.5$ and $b = 1$. Figure 4.9 shows that for overdamped ($\zeta > 1$), minimum-phase ($b \geq 0$) system, unbiased delay estimation is guaranteed if the sampling time satisfies Equation 4.47. The condition given on the sampling rate is not necessary but rather a sufficient one, as can be easily seen in Figure 4.9.

Figure 4.9 shows that unbiased delay estimation is also achievable for under-damped systems. If the system has a minimum-phase zero ($b > 0$), then any sampling rate is adequate for unbiased delay estimation regardless of the system damping ratio. If the system has no zero ($b = 0$), then the sampling rate required for unbiased delay estimation is still in the range defined by Equation 4.47 for $\zeta > .5$. Systems having $\zeta < .5$ are not very common in industry and for these systems the sampling rate requirements imposed by the FMVRE is usually slower than that required for estimating the other parameters.

Example 4.2

Figure 4.10 shows the delay estimation for the following parameter variation:

<table>
<thead>
<tr>
<th>Time</th>
<th>$T_d$</th>
<th>$\zeta$</th>
<th>$b$</th>
<th>$w_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 &lt; t &lt; 100$</td>
<td>7</td>
<td>1.0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$100 &lt; t &lt; 200$</td>
<td>4</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$200 &lt; t &lt; 300$</td>
<td>1</td>
<td>1.0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE)

$b = 0$

$b = 0.5$

$b = 1$

Figure 4.9: Minimum $T$, required for a second order system.
The sampling time is one second which satisfies Equation 4.47 for the range of parameter variations. Figure 4.10 also shows that the condition of white excitation is a sufficient one because the delay is correctly estimated using a square wave input. The controller used is the no zero cancellation pole-placement given in Chapter 6. The delay estimation converges in less than 20 samples.

4.5 Higher-Order Systems

Systems of order higher than two have no general form to study and no exact condition for unbiased delay estimation is achievable. However, high-order systems in industry are often over damped. In practice, these high-order over-damped systems are approximated by a low-order-plus-delay model. Our goal now is not to have an unbiased delay estimate of the system but rather to have a biased delay estimate that well approximates the high-order dynamics and the FMVRE automatically gives this approximation.

Consider the system:

$$G(s) = \frac{1}{(s + 1)^8}$$

which has no time delay \((k = 0)\). The system has the step response shown in Figure 4.11. The sampling time is selected to be \(T_s = 1\)sec.. For this sampling time, Condition 4.2 is not satisfied because, as shown in Figure 4.11, the maximum of the function \(E_1\) occurs at the fifth sample \((k = 4)\) instead of the first sample. However, using the Ziegler-Nichols method we can obtain a good system approximation as a low-order system plus four seconds time delay \((k = 4)\). This is exactly the value obtained from the FMVRE method.

On the other hand, for white input, the standard cross-correlation between the input and the output will have the relation shown in Figure 4.11 which is also the impulse
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE)

Figure 4.10: FMVRE of a second order system.
Figure 4.11: Response of a high order system.
response of the system. The system delay equals the time to the first non-zero sample of the impulse response. In practice, the limit between what is considered zero and what is considered non-zero highly depends on the particular system under study. In our case, the limit is over 50% of the maximum value in order to obtain the same good approximation as by maximizing $E_1$. It is clear that looking for the maximum of the function $E_1$ is more practical than looking for the first non-zero of the cross-correlations function. The simulation results for this system are shown in Figure 4.12. The delay estimation converges in less than 30 samples.

All of the preceding analysis was for noise-free systems. However, in open loop, all the results and conclusions are still valid even in the presence of coloured random noise as shown in the following analysis.

4.6 Effect of Random Noise

**Theorem 4.2** The estimated delay ($\hat{k}$) that maximizes the function $E_1$ is unbiased in presence of process noise that is uncorrelated with the system input.

Proof: Assume that the Gaussian random white noise $e(t) = N(0, \sigma)$ affects the measurements through the stable filter $G_e$, i.e.:

$$y(t) = G(q^{-1})u(t) + G_e(q^{-1})e(t)$$

$$= y_u + y_e$$

where $y_u$ is the output due to the input and $y_e$ is the output due to the noise. The input-output cross-correlation is given by:

$$r_{uy}(t) = r_{uy_u}(t) + r_{uy_e}(t)$$

The term $r_{uy_u}$ represents the deterministic part of the cross-correlation function and is given by Equation 4.34. To solve for the stochastic part we first expand $G_e(q^{-1})$ as an
Figure 4.12: FMVRE of a high-order system.
infinite moving average polynomial similar to that of $G(z)$:

$$G_e(q^{-1}) = w_0 + w_1 q^{-1} + w_2 q^{-2} + \cdots$$  \hfill (4.51)

that means:

$$y_e(t) = w_0 e(t) + w_1 e(t - 1) + w_2 e(t - 2) + \cdots$$  \hfill (4.52)

then the stochastic part of the cross-correlation function is given by:

$$r_{wy_e}(t) = w_0 r_{ue}(t) + w_1 r_{ue}(t + 1) + w_2 r_{ue}(t + 2) + \cdots$$  \hfill (4.53)

since $e$ and $u$ are uncorrelated. This means that the noise has no effect on the cross-correlation function and consequently it has no effect on the function $E_1$ and this proves the theorem. \hfill $\Box$

This result means that the presence of coloured noise does not affect $E_1$ in any way, even by increasing or decreasing $E_1$. In the EMVRE where we minimize $J$, we showed that the coloured noise will increase $J$ and so for small signal to noise ratio the maximum of $J$ will not be sharp. While in the FMVRE method, the signal to noise ratio does not affect $E_1$ and it is only the excitation richness that determines $E_1$ for a specific system.

The result of this theorem is very important because it guarantees unbiased delay estimation under the same conditions that will surely result in a biased delay estimation using least-squares estimator and the extended polynomial method. In fact with biased estimates, the extended polynomial method fails to determine the delay because none of the extended polynomial parameters will converge to zero.

Example 4.3
Consider the system:

\[ y(t+1) = 0.82y(t) + 0.18u(t-3) + e(t+1) - 2e(t) + e(t-1) \]

where the input is the square wave shown in Figure 4.13. The noise sequence \( e(t) \) is a random signal ranging between \(-0.5\) and \(0.5\) with equal probability. The delay estimation converges to the correct value. If the parameters of this system are to be estimated, the Approximate Maximum Likelihood estimation method can be used in a variable regression mode.

### 4.7 Excitation Requirement for FMVRE

The preceding analysis of the FMVRE was based on a white excitation. The autocorrelation function for the white excitation is an impulse function. In this section we consider an excitation signal with a more general autocorrelation function.

In order to get good parameter estimates, the excitation signal must be rich enough. The number of different frequency components in the excitation signal should be at least half the number of unknown parameters. For example, for a first-order system having unknown gain and pole location, one sinusoid is a sufficient excitation provided that its frequency is close to the system corner frequency. If the system has an unknown delay, then a richer excitation is required if either the extended numerator polynomial or the rational approximation methods is used. In this section we prove that for such a system, the FMVRE requires only one sinusoid to estimate both the model parameters and the delay.

Consider the input \( u(t) \) that has autocorrelation \( r_u(\tau) \). The cross-correlation between the system input and output is given by:

\[
r_{uy}(\tau) = \sum_{i=0}^{\infty} y_{impulse}(i) r_u(\tau - i)
\]  (4.54)
Figure 4.13: FMVRE in presence of coloured noise.
The convolution expression in the last equation implies that the cross-correlation is generated from the auto correlation in the same way as the generation of the system output from the input. The function $E_1$ is then generated from the input auto-correlation through the filter $(1 - z^{-1})G(z)$.

In practice, a periodic signal is commonly used as the excitation signal, for instance, a PRBS, a square wave, a sinusoidal wave, etc. In this case, the function $E_1$ is also periodic with the same frequency and its maximum appears repetitively. In order to obtain a unique value for the estimated delay, there must be only one maximum of $E_1$ in the range $[k_{\text{min}}, k_{\text{max}}]$. This leads to the lower bound on the excitation signal period $T_e$:

$$T_e > T_s(k_{\text{max}} - k_{\text{min}})$$  \hspace{1cm} (4.55)

On the other hand, the frequency of the excitation signal should be fast enough to produce a sensible phase difference between the input and the output. An upper bound on a sinusoidal excitation period is derived in the next section.

**4.7.1 Sinusoidal Excitation**

**Theorem 4.3** Using the FMVRE, the estimated delay for a minimum phase stable system that has a pole-excess of one is unbiased if the sinusoidal excitation period is related to the fastest time constant of the system $\tau_{\text{min}}$ by:

$$T_e \leq \pi \sqrt{2T_s\tau_{\text{min}}}$$  \hspace{1cm} (4.56)

Proof: The FMVRE does not impose constraints on the sampling rate choice. Then, the sampling time $T_s$ is independent of the delay estimation and is usually selected to be faster than the fastest time constant of the system to provide good parameter
estimation and closed loop control. This fast sampling is required when assuming sinusoidal input and output.

Now, consider a stable minimum phase system of pole excess equal to one. If the excitation frequency $\omega_e$ is higher than the highest corner frequency (corresponding to $\tau_{\text{min}}$), then the phase shift of the system is approximated by:

$$\arg(G(z)) \approx -(\omega_e T_d - (\pi/2 - \frac{1}{\omega_e \tau_{\text{min}}})$$

The first term in the last equation represents the phase shift due to the system time delay. The second term approximates the phase shift due to the other system dynamics. The excitation signal is sinusoidal and so its auto-correlation function is:

$$r_u(\tau) = r_u(0) \cos(\omega_e \tau)$$

This function has an amplitude of $r_u(0)$ and is taken as our reference sinusoid, so its phase is zero. Due to the existence of the sample and hold in the system, the sampled auto-correlation function will have a phase of one half of a sample, i.e.:

$$\arg(r_u) = -\omega_e T_s/2$$

It is then direct to show that:

$$\arg((1 - z^{-1})r_u) = \pi/2 - \omega_e T_s$$

The phase of the function $E_1$ is then given by:

$$\arg(E_1) = \arg(G(1 - z^{-1})r_u) = -\omega_e T_d - \omega_e T_s + \frac{1}{\omega_e \tau_{\text{min}}}$$

For correct delay estimation, the maximum of $E_1$ should occur at the $(\hat{k} + 1)$ th sample, which corresponds to a phase shift of $\arg(E_1) = -\omega_e(\hat{k} + 1)T_s = -\omega_e(T_d + T_s)$. It is
then necessary that the last term in the last equation contributes a phase shift of less than one half of a sample so that the sample at which $E_1$ is maximum does not change. The condition is then:

$$\frac{1}{\omega_e \tau_{\text{min}}} \leq \omega_e T_s / 2$$

and the excitation frequency is given by:

$$\omega_e \geq \sqrt{\frac{2}{T_s \tau_{\text{min}}}}$$

and since

$$\omega_e = \frac{2\pi}{T_e}$$

then

$$T_e \leq \pi \sqrt{2T_s \tau_{\text{min}}}$$

and this proves the theorem. $\square$

Computer simulations show that when the upper bound of the excitation signal given here is exceeded, a correct delay estimation can still be obtained. We believe that the reason is that the sampling process of the sinusoid does not only result in a phase shift but also in a distortion of the wave. The sampled sinusoid actually has more than one frequency component which means a richer excitation. Consequently, the condition based on one sinusoid can be relaxed.

**Example 4.4**

This simulation example demonstrates the results of Theorem 4.3. Consider the first order system:

$$G(s) = e^{-T_s s} \frac{1}{\tau s + 1}$$

(4.57)

The system delay and time constant change with time as follows:
The system is stable and minimum phase. For this system: $T_e = 1$ sec, $k_{max} = 4$, $k_{min} = 0$ and $\tau_{min} = 3$ sec is the fastest expected time constant. The conditions for the excitation frequency are then:

$$T_e > (4 - 0) = 4 \text{ sec}$$

$$T_e \leq \pi \sqrt{(2)(3)(1)} = 7.69 \text{ sec}$$

The value of $T_e = 6$ sec is selected which satisfies the excitation requirement. The parameters and the delay of the corresponding discrete model are shown in Figure 4.14.

The results show that one sinusoid is a sufficient excitation to identify the system.

### 4.8 Robustness of the FMVRE

The approximation of a high-order system by a low-order model usually suffers from lack of robustness. If a stable system is excited by a high frequency and the resulting system phase cannot be described by the low-order model, then the estimated model can become unstable and this can destabilize the closed-loop. The following example illustrates this fact (Åström 1984-c).

**Example 4.5**

Consider a model given by:

$$G_m(s) = \frac{\dot{b}}{s + \dot{a}}$$

(4.58)
Figure 4.14: FMVRE using only one sinusoid.
used to describe the system:

\[ G_s(s) = \frac{458}{(s + 1)(s^2 + 30s + 229)} \]  

(4.59)

The system dynamics correspond to the nominal plant \( \frac{2}{s+1} \) cascaded with \( \frac{229}{s^2 + 30s + 229} \). The system thus has two poles \( s = -15 \pm 2j \), which have been neglected in the model used. If the excitation signal is sinusoidal of frequency \( \omega \) then in order to match the system response at this frequency, the model should satisfy:

\[ \frac{\hat{b}}{j\omega + \hat{a}} = \frac{458/(259 - \omega^2)}{j\omega + (229 - 31\omega^2)/(259 - \omega^2)} \]  

(4.60)

then the model parameters are given by:

\[ \hat{b} = \frac{458}{259 - \omega^2} \]
\[ \hat{a} = \frac{229 - 31\omega^2}{259 - \omega^2} \]

The model describes the system accurately only at very small \( \omega \). As \( \omega \) increases, we get the following estimates:

- \( 0 < \omega < 2.71 \) \( \hat{a} > 0 \) stable model
- \( 2.71 \leq \omega < 16.1 \) \( \hat{a} \leq 0 \) unstable model
- \( \omega = 16.1 \) \( \hat{a} = \pm \infty \) unstable model
- \( \omega > 16.1 \) \( \hat{a} > 0 \) stable model

It is clear that the presence of unmodelled dynamics can destabilize the controller if the excitation frequency is high enough.

Introducing a variable delay in the low-order model rectifies this problem. The phase of the delay part in the model is the product of the exciting frequency and the modelled time delay, and this phase can be as large as required to match the phase of the system. The resulting model is always stable and exactly matches the system response at the exciting frequency as proven in the following theorem.
Theorem 4.4 A model of the form:

\[ G_m(s) = e^{-T_d s} \frac{b}{s + a}, \]

\[ G_m(z) = z^{-k} \frac{b_0}{z + a_1} \]

where \( \hat{T}_d \) is the time delay estimated by the FMVRE, is stable and exactly matches the response of any stable system having a negative phase at a given frequency.

Proof: Let the magnitude and phase of the system at the excitation frequency \( \omega_e \) be given by \( |G_e(j\omega)| \) and \( \phi_e \) respectively. From the sinusoidal excitation analysis, the phase of the function \( E_1 \) is given by:

\[ \arg(E_1) = \arg(G(1 - z^{-1})r_u) \]

\[ = \frac{\pi}{2} - \omega_e T_s + \phi_s \]  

(4.61)

Since the maximum of \( E_1 \) occurs at the \((k + 1)\) sample corresponding to the shift \( \arg(E_1) = -\omega_e(k + 1)T_s = -\omega_e(\hat{T}_d + T_s) \), then \( \hat{T}_d \) is given by:

\[ -\omega_e \hat{T}_d = \frac{\pi}{2} + \phi_s \]

but since \( \hat{T}_d \) is an integer multiple of \( T_s \), the exact expression of \( \hat{T}_d \) is:

\[ \hat{T}_d = T_s \text{integer}\left(\frac{\pi/2 + \phi_s}{\omega_e T_s}\right) \]  

(4.62)

then,

\[ \frac{\pi}{2} + \phi_s - T_s \omega_e \leq -\omega_e \hat{T}_d < \frac{\pi}{2} + \phi_s \]  

(4.63)

Now let the phase resulting from the first-order part of the model be \( \phi_m \), then the estimation of the model parameters should satisfy:

\[ \phi_m - \omega_e \hat{T}_d = \phi_s \]  

(4.64)
Substituting the last equation in the one before we get:

$$-\pi/2 < \phi_m \leq -\pi/2 + T_s \omega_e$$

which means that:

$$0 < \hat{a}/\omega_e \leq \tan(T_s \omega_e)$$

So \(\hat{a}\) is always positive which means a stable model that matches the phase of the system. To match the magnitude of the system and the model, \(\hat{b}\) must satisfy:

$$\hat{b} = |G_s(j\omega_e)|\sqrt{\omega_e^2 + a^2}$$

which is always positive and finite and this completes the proof. □

It is to be noted that the estimated delay variation required for exact frequency response matching is not large. At low frequency the phase of the system is \(|\phi_s| < (\pi/2 - \omega_e T_s)\) and this phase is totally represented by the first-order model phase, then the estimated delay is zero \((\hat{T_d} = 0, \hat{k} = 0)\). At high frequency the phase of the system reaches its final value equal to the system pole excess times \(\pi/2\) and the expression in Equation 4.62 will reach the value of zero. Between these two frequency extremes, \(\hat{k}\) will reach a maximum given by the ratio in Equation 4.62.

Example 4.6

Now let us apply Theorem 4.4 to the system in Example 4.5. The Bode plot of the system is shown in Figure 4.15. The operating frequency is selected to be \(\omega_e = 16.1 \text{rad/sec}\) which corresponds to a phase shift of \(180^\circ\)\(\phi_s = -\pi\). Example 4.5 shows that the parameters of the first order model are both infinitely large at this frequency.

We start by selecting a sampling frequency adequate for the operating frequency, and this is taken as 10 times the operating frequency. This means that the phase shift
Figure 4.15: Bode plot of the third-order system
due to one sample is $\omega_e T_s = \pi/5$ and $T_s = 0.039\text{sec}$. According to Equation 4.63, the estimated delay should be given by:

$$
\frac{(\pi/2) - (\pi) - (\pi/5)}{T_s} \leq -(\omega_e T_s)(\hat{T}_d/T_s) < \frac{\pi}{2} - (\pi)
$$

$$
-\frac{7}{10}\pi \leq -(\pi/5)(\hat{T}_d/T_s) < -\pi/2
$$

then:

$$
-3.5 \leq -(\hat{T}_d/T_s) < -2.5
$$

and since $\hat{k} = \hat{T}_d/T_s = \text{integer}$, then $\hat{k} = 3$ is the only value possible. Figure 4.16 shows the excitation signal and the estimated delay that converges to the expected value of 3. The phase shift due to the model pole is given by Equation 4.65 as:

$$
\phi_m = \phi_s + \omega_e \hat{T}_d = \phi_s + \omega_e T_s(\hat{T}_d/T_s)
$$

$$
= -\pi + 3\pi/5 = -2\pi/5
$$

$$
= -\tan(\omega_e/\hat{a})
$$

and hence the discrete model pole is given by:

$$
\hat{a}_1 = -\exp(-T_s \hat{a}) = -\exp((\omega_e T_s)/(-\omega_e/\hat{a}))
$$

$$
= -\exp((\omega_e T_s)/(-\tan^{-1}(\phi_m))) = \exp((\pi/5)/(-\tan^{-1}(-2\pi/5)))
$$

$$
= -.815
$$

which is exactly the value estimated, as shown in Figure 4.17.

This example showed the improvement in the model robustness when a delay is included in the model and the operating frequency is a sufficient excitation to estimate the delay.
Figure 4.16: Delay estimation for the case of model order mismatch
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE)

Figure 4.17: Parameter estimation for the case of model order mismatch
4.9 Systems with fractional delay

By using the standard $Z$ transform in all of the preceding analysis, we assume that the system delay ($T_d$) is an integer multiple of the sampling time ($T_s$). If the system has a fractional delay, the *modified $Z$ transform* ($Z_m$) is used (Kuo, 1980). The zeros of the sampled system change significantly if fractional delay exists, while the poles of the sampled system do not change. For example, consider the first-order system:

$$Y(s) = e^{-T_d s} \frac{1}{\tau s + 1} U(s) \quad (4.70)$$

where:

$$T_d = k T_s + (1 - m) T_s, \quad 0 < m \leq 1$$

the system delay has $k$ integer number of samples plus a fraction of a sample determined by the value of $m$. The sampled system is given by:

$$y(t) = q^{-k} \frac{b_0 q^{-1} + b_1 q^{-2}}{1 + a_1 q^{-1}} u(t) \quad (4.71)$$

where:

$$a_1 = -e^{-T_s/\tau}$$
$$b_0 = 1 - e^{-mT_s/\tau}$$
$$b_1 = e^{-mT_s/\tau} - e^{-T_s/\tau}$$

If $m = 1$, then $b_1 = 0$ and $b_0 = 1 + a_1$ which corresponds to the sampled system with $k$ samples of delay. In this case, there is no zero in the sampled model. For other values of $m$ there will be a zero in the sampled model given by the ratio $b_1/b_0$. This zero can easily be outside the unit circle, which means a non-minimum phase model (Åström et al. 1984-b). In general, the presence of fractional delay can be modelled by increasing
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE)

the number of zeros in the model by one. In other words, the presence of fractional delay means that \( \text{deg} B = n + 1 \).

In the FMVRE we maximize the function \( E_1(\hat{k}) = r_{uv}(1 + \hat{k}) \). If \( \hat{k} = k_{m=1} \) is the value that maximizes \( E_1 \) for integer delay, then \( \hat{k} = k_{m=1} + (1 - m) \) is the value that maximizes \( E_1 \) for fractional delay. But because the values of \( E_1 \) are considered only at integer values of \( \hat{k} \), then \( E_1 \) is maximum at either \( \hat{k} = k_{m=1} \) or at \( \hat{k} = k_{m=1} + 1 \) depending on the system under consideration and the value of \( m \). This means that the estimated delay can be one step more than the value estimated for integer delay. This difficulty can be overcome by extending the numerator polynomial by one extra parameter. Let the model for integer delay be:

\[
G_{m=1}(z) = z^{-k_{m=1}} \frac{\hat{b}_0 z^{-1} + \cdots + \hat{b}_{n-1} z^{-n}}{1 + \hat{a}_1 z^{-1} + \cdots + \hat{a}_n z^{-n}}
\]

then, the model for fractional delay should be:

\[
G_{m\neq1}(z) = z^{-\hat{k}} \frac{\hat{b}_{-1} + \hat{b}_0 z^{-1} + \cdots + \hat{b}_{n-1} z^{-n} + \hat{b}_n z^{-n-1}}{1 + \hat{a}_1 z^{-1} + \cdots + \hat{a}_n z^{-n}}
\]

Two parameters are added to the numerator polynomial of \( G_{m=1} \); the last parameter in the polynomial (\( \hat{b}_n \)) represents the increase in the number of zeros and the first parameter (\( \hat{b}_{-1} \)) corrects the bias that may result in the delay estimation. If a bias results in the delay estimation (\( \hat{k} = k_{m=1} + 1 \)), then \( \hat{b}_n = 0 \) and \( \hat{b}_{-1} \neq 0 \) give the required model. On the other hand, if \( \hat{k} = k_{m=1} \), then \( \hat{b}_n \neq 0 \) and \( \hat{b}_{-1} = 0 \) give the required model.

Not every system will require this extension of the numerator polynomial because of the delay estimation bias. A first-order systems with rich excitation will have the function \( r_{uv} \) similar to the system impulse response. This means that \( E_1(\hat{k}) \) is zero for \( \hat{k} < k \) and \( E_1(\hat{k}) \) is maximum for \( \hat{k} = k \) and \( 0 < m \leq 1 \). The presence of fractional delay does not affect the delay estimation for first order system with rich excitation.
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE)  

The following example demonstrates the FMVRE+RLS estimation of the model in presence of a fractional delay.

Example 4.7

Consider the first-order system with fractional delay:

\[
G(s) = e^{-T_d s} \frac{1}{\tau s + 1},
\]

\[
G(z) = z^{-k} \frac{b_0 z^{-1} + b_1 z^{-2}}{1 + a_1 z^{-1}}
\]

The sampling time is \( T_s = 1 \text{sec.} \) and system time constant \( \tau = 4 \text{sec.} \). The time delay \((T_d)\) variation and the resulting sampled system parameter are as follows:

<table>
<thead>
<tr>
<th>( t )</th>
<th>( T_d )</th>
<th>( m )</th>
<th>( k )</th>
<th>( a_1 )</th>
<th>( b_0 )</th>
<th>( b_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>([0, 150])</td>
<td>1.5</td>
<td>.5</td>
<td>1</td>
<td>.7788</td>
<td>.1175</td>
<td>.1036</td>
</tr>
<tr>
<td>([150, 300])</td>
<td>2.75</td>
<td>.25</td>
<td>2</td>
<td>.7788</td>
<td>.0605</td>
<td>.1606</td>
</tr>
<tr>
<td>([300, 450])</td>
<td>3.5</td>
<td>.5</td>
<td>3</td>
<td>.7788</td>
<td>.1175</td>
<td>.1036</td>
</tr>
<tr>
<td>([450, 600])</td>
<td>4.25</td>
<td>.75</td>
<td>4</td>
<td>.7788</td>
<td>.1709</td>
<td>.0502</td>
</tr>
<tr>
<td>([600, 750])</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>.7788</td>
<td>.2212</td>
<td>0</td>
</tr>
</tbody>
</table>

The input used is a square wave of 40 samples period. The delay estimation is shown in Figure 4.18. The delay estimation converges to its correct value for the different value of \( m \), which means that the square wave input is a rich excitation for the system and no need to extend the numerator polynomial. Figure 4.19 shows the estimates of the model parameters \( a_1, b_0 \) and \( b_1 \). The model parameters converge to their correct values for the different values of \( m \).
Figure 4.18: Delay estimation for system with fractional delay
Chapter 4. The Fixed Model Variable Regression Estimation (FMVRE)

Figure 4.19: Parameter estimation for system with fractional delay
4.10 Conclusions

In this chapter, a fixed model is used with the VRE, the result is the FMVRE. The FMVRE has the same advantages over the extended polynomial method and the rational approximation method as the EMVRE since no additional parameters are required to estimate the delay. However, the FMVRE does have unique properties that make it superior to the EMVRE. These advantages are:

1. The delay estimation is decoupled from the parameter estimation and consequently it always converges for first and second-order systems.
2. The delay estimation is bounded for any sampling rate.
3. The delay estimation for a high-order over-damped system is equivalent to the delay obtained by the Ziegler-Nichols method. The method can therefore be used for the auto-tuning of controllers based on the Ziegler-Nichols approximation.
4. The delay estimation does not increase the excitation richness requirements over those for good estimation of the system parameters.
5. The estimated model is stable in the presence of unmodelled dynamics and exactly matches the frequency response of the system at the operating frequency.
Chapter 5

VRE-Based Adaptive Control

In this chapter the model parameters estimated by the RLS and the model delay estimated using the VRE are used in the controller design and the response is compared with that obtained if the design is based on the extended numerator polynomial method. Three controllers are covered in this chapter and other controllers can also be used. The controllers selected here are the Pole-Placement, the Dahlin and the Generalized Predictive controllers. They are chosen because of their practical importance and because of their capability of regulating systems with dominant time delay (Åström 1989, Dumont 1982-b and Clarke 1988). The design is based on the certainty equivalence principle which assumes exact knowledge of the system model. Then the estimated model is used in the design instead of the actual one. In all cases, only deterministic systems are considered. Stochastic systems can be dealt with systematically as extensions to the designs given here. The controller design is indirect which means that the estimated parameters are not those of the controller. Direct controller design is also possible. The structure of the estimators and the controller is shown in Figure 5.20.

The three adaptive controllers will be applied to the following system:

$$G(s) = e^{-T_s s} \frac{bs + 1}{s^2/w_o^2 + 2\zeta s/w_o + 1}$$  \hspace{1cm} (5.72)

The sampling rate is $T_s = 1$ sec and the system parameters and delay change with time as follows:
Figure 5.20: Structure of the Estimators and the adaptive controller
Time | $T_d$ | $\zeta$ | $b$ | $w_0$
---|---|---|---|---
$0 < t < 100$ | 7 | 1.0 | 0 | 1
$100 < t < 200$ | 4 | 1.5 | 1 | 1
$200 < t < 300$ | 1 | 2.0 | 1 | 1
$300 < t < 400$ | 1 | 2.0 | -1 | 1

The input to the system is a square wave of a 20 samples period. The values of $k_{\min} = 0$ and $k_{\max} = 8$ are used in the delay estimation. The system is non-minimum phase for $t \in [300, 400]$ because $b$ is negative which means that special precautions should be taken in the control design (Åström 1980-b and Clarke 1984). The non-minimum phase system can be written as:

$$G(s) = G_{NMP} G_{MP}$$

where:

$$G_{NMP} = \frac{-|b|s + 1}{|b|s + 1}$$

$$G_{MP} = e^{-T_d s} \frac{|b|s + 1}{s^2/w_0^2 + 2\zeta s/w_o + 1}$$

The second-order system analysis of the last chapter showed that the pure delay term in $G_{MP}$ can be accurately estimated using the FMVRE. The rational function $G_{NMP}$ has a unity gain at all frequencies, the same as a pure delay function. The phase of $G_{NMP}$ is the same as that of the pure delay function $e^{(-2|b|s)}$ at low frequencies. In fact, $G_{NMP}$ is found as a fundamental building block in many delay rational function approximations, e.g. Padé, Laguerre and all-pass filters. This means that at low frequencies ($\omega |b| < 1$) the non-minimum phase system can be approximated by a minimum phase model having delay. At higher frequencies the approximation can still be used but with a lesser delay to approximate $G_{NMP}$. The FMVRE can automatically approximate
The open loop FMVRE is shown in Figure 5.21 where the estimated delay converges to the value of 2 which agrees with the above analysis.

5.1 Adaptive Pole-Placement

In a Pole-Placement controller (PP) (Åström and Wittenmark 1980-a and Wellstead et al. 1979), the input-output ARMA model is used:

\[ A(q^{-1})y(t) = q^{-k}B(q^{-1})u(t) \]  
(5.76)

Let the desired response from the reference signal \( u_c \) to the output be described by the dynamics:

\[ A_m(q^{-1})y(t) = q^{-k}B_m(q^{-1})u_c(t) \]  
(5.77)

A general linear controller can be described by:

\[ R(q^{-1})u(t) = T(q^{-1})u_c(t) - S(q^{-1})y(t) \]  
(5.78)
Figure 5.21: Open-loop response and FMVRE delay estimation
Elimination of $u$ between Equation 5.76 and 5.78 gives:

$$y = q^{-k} \frac{BT}{AR + q^{-k}BS} u_c$$  \hspace{1cm} (5.79)

To achieve the desired input-output relation, the following condition must hold:

$$\frac{BT}{AR + q^{-k}BS} = \frac{B_m}{A_m}$$  \hspace{1cm} (5.80)

The denominator $AR + q^{-k}BS$ is the closed-loop characteristic polynomial. To carry out the design, the polynomial $B$ is factored as:

$$B = B^+ B^-$$  \hspace{1cm} (5.81)

where $B^+$ is a monic polynomial whose zeros are stable and so well damped that they can be cancelled by the regulator. When $B^+ = 1$, there is no cancellation of any zero. Since $B^+$ is cancelled, it also factors the closed-loop characteristic polynomial. The other factors of this are $A_m$ and $A_0$, where $A_0$ is called the observer polynomial. $A_0$ must be stable and its dynamics does not appear in the output. This gives the Diophantine equation:

$$AR + q^{-k}BS = A_0 A_m B^+$$  \hspace{1cm} (5.82)

It follows from this equation that $B^+$ divides $R$. Hence:

$$R = R_1 B^+$$  \hspace{1cm} (5.83)

$$AR_1 + q^{-k}B^-S = A_0 A_m$$  \hspace{1cm} (5.84)

The solution of the Diophantine equation is essentially the same as solving a set of linear equations. The solution is unique if $A$ and $B^-$ are relatively prime. It follows from Equation 5.80 that $B^-$ must divide $B_m$ and that:

$$T = A_0 B_m / B^-$$  \hspace{1cm} (5.85)
Chapter 5. VRE-Based Adaptive Control

The pole placement design procedure can be summarized as follow:

given $A, B, k, A_m, B_m, A_0$,

Step 1: Factor $B$ as $B = B^+B^-$

Step 2: Solve $R_1$ and $S$ from the equation $AR_1 + q^{-k}B^-S = A_0A_m$.

Step 3: Form $R = R_1B^+$ and $T = A_0B_m/B^-$. The control law is given by:

$$Ru = Tu_c - Sy$$

To combine the pole placement design with the parameter and delay estimation, we simply replace $A, B, k$ in the above design by their estimates $\hat{A}, \hat{B}, \hat{k}$ respectively. Because the polynomial $\hat{B}$ is not known a-priori, it is safe not to cancel any of estimated model zeros, which means that $B^+ = 1, B^- = B$. If there is no delay estimation and the extended polynomial method is used instead of the FMVRE, the above design is still valid but $k$ is replaced by $k_{\text{min}}$ and $\text{deg}B$ is $n + k_{\text{max}} - k_{\text{min}}$. This means that $\text{deg}R_1 = n + k_{\text{max}}$ instead of $\text{deg}R_1 = n + \hat{k}$ if the FMVRE is used. The reduction of the number of controller parameters directly reduces the number of equations to be solved and this means much less computations performed on-line.

The adaptive pole placement is now applied to the system in Equation 5.72. The model used is second-order and the controller has no zero cancellation. The desired closed-loop characteristic equation has one pole corresponding to a time-constant of one sample and all other poles are at the origin of the unit circle.

The adaptive PP/FMVRE response is shown in Figure 5.22 where the model delay converges to the actual system delay for the minimum-phase cases and to the value of 2 for the non-minimum phase case. The closed-loop response compares closely with the reference signal for both the minimum phase and the non-minimum phase cases. The model parameter estimates are shown in Figure 5.23. The delay estimation converges
in less than 20 samples while the parameter estimation converges much slower. It is expected then that when the extended numerator polynomial method is used instead of the VRE, the convergence is slow and thus the closed-loop performance is poor. This is shown in Figure 5.24 where the adaptive PP is designed based on the extended numerator polynomial method. Giving enough time for the model parameters to converge to their final values, the response using the extended numerator polynomial will be the same as that using the FMVRE except for the non-minimum phase system part where the extended numerator polynomial will eventually result in a better closed-loop response because it represents the system exactly. The main advantage of using the FMVRE in the design is the speed of getting a good closed-loop response. The EMVRE based adaptive PP also has the advantage of obtaining a good closed-loop response quickly. This is shown in Figure 5.25. The response of the adaptive PP/EMVRE may be different from that of the adaptive PP/FMVRE in the delay learnings period which is usually the first 20 samples after the delay changes. After the delay learning period, the adaptive PP/EMVRE gives the same good closed-loop response as the adaptive PP/FMVRE.

5.2 Adaptive Dahlin Controller

The Dahlin controller was originally proposed by Dahlin (1968). It can be thought of as a special case of the general pole placement design of the previous section. The closed-loop response required is that of a first-order with a gain of one and with a time delay equal to the open-loop delay. Also the controller input is the error between the actual output \(y\) and the desired one \(u_c\), this means that \(T = S\) in the general pole placement given above.
Figure 5.22: Adaptive PP/FMVRE response and delay estimation
Figure 5.23: Adaptive PP/FMVRE parameter estimates
Figure 5.24: Adaptive PP/extended polynomial response
Figure 5.25: Adaptive PP/EMVRE response and delay estimation
Let $Q$ be the required closed loop pole, then:

$$A_m(q^{-1}) = 1 - Qq^{-1}$$
$$B_m(q^{-1}) = B^-(q^{-1})A_m(1)/B^-(1)$$

The required closed loop response is given by:

$$y(t) = q^{-k} \frac{B_m}{A_m} u_c(t) = q^{-k} \frac{B^-}{B^-(1)} \frac{1 - Q}{1 - Qq^{-1}} u_c(t) \tag{5.86}$$

The controller is then given by:

$$u(t) = \frac{T}{R}(u_c(t) - y(t)) = A \frac{B_m}{B A_m - q^{-k} B_m} (u_c(t) - y(t))$$
$$= \frac{A}{B^-(1)B^+(A_m/A_m(1) - q^{-k}B^-/B^-(1))} (u_c(t) - y(t)) \tag{5.88}$$

The controller automatically has an integrator because $R(1) = 0$ and this property of the controller is required to compensate for set point changes.

The adaptive Dahlin controller is applied to the system in Equation 5.72 with the same closed-loop requirements as in the adaptive pole placement design. The response of the adaptive Dahlin/FMVRE is shown in Figure 5.26. The closed-loop response is good for both the minimum-phase and the non-minimum phase system. The delay estimate converges in less than 20 samples and the model parameters converge much slower than that as shown in Figure 5.27. The response of the adaptive Dahlin controller based on the extended numerator polynomial method is shown in Figure 5.28 which shows a slow closed-loop response. The response of the adaptive Dahlin/EMVRE is also much better than that of the extended numerator polynomial design as shown in Figure 5.29. The use of VRE-based design accelerates the achievement of a good closed loop response.
Figure 5.26: Adaptive Dahlin/FMVRE response and delay estimation
Figure 5.27: Adaptive Dahlin/FMVRE parameter estimates
Figure 5.28: Adaptive Dahlin/extended polynomial response
Figure 5.29: Adaptive Dahlin/EMVRE response and delay estimation
5.3 Adaptive Generalized Predictive controller

In the Generalized Predictive Control (GPC), the objective of the controller is to minimize the following loss function (Clarke et al. 1987-a,b):

\[
J(N_1, N_2, N_u) = E\left\{ \sum_{d=N_1}^{d=N_2} (y(t+d) - y_m(t+d))^2 + \sum_{d=1}^{d=N_u} \rho \Delta u(t + d - 1)^2 \right\} \tag{5.89}
\]

Different choices of \( N_1, N_2, N_u \) and \( \rho \) give rise to different schemes suggested in the literature. \( y_m(t + d) \) is the required future output \( d \) steps ahead and \( \rho \) is the control weight. The future predictions in the loss function are estimated from the prediction model:

\[
y(t+d) = F_d B \Delta u(t + d - k - 1) + G_d y(t) \tag{5.90}
\]

where \( F_d \) and \( G_d \) are computed from the identity:

\[
1 = A(q^{-1})F_d(q^{-1})(1 - q^{-1}) + q^{-d}G_d(q^{-1}) \tag{5.91}
\]

The future outputs can be written as:

\[
\begin{align*}
y(t+1) &= R_1 \Delta u(t-k) + \bar{y}_1(t) \\
y(t+2) &= R_2 \Delta u(t+1-k) + \bar{y}_2(t) \\
&\vdots \tag{5.92} \\
y(t+N) &= R_N \Delta u(t+N-1-k) + \bar{y}_N(t)
\end{align*}
\]

where \( R_d \) and \( \bar{y}_d \) are given by:

\[
B(q^{-1})F_d(q^{-1}) = R_d(q^{-1}) + q^{d-k}\bar{R}_d(q^{-1}) \tag{5.93}
\]

\[
\bar{y}_d(t) = \bar{R}_d(q^{-1})u(t) + G_d(q^{-1})y(t) \tag{5.94}
\]

The variable \( \bar{y}_d(t) \) can be interpreted as the constrained prediction of \( y(t+d) \), under the assumption that present and future control signals are zero.
Equation 5.92 can be written in the following vector form:

\[
y = R \Delta u + \bar{y}
\]

where:

\[
y = \begin{bmatrix} y(t+1) & \cdots & y(t+N) \end{bmatrix}^T
\]

\[
\Delta u = \begin{bmatrix} \Delta u(t+1-k) & \cdots & \Delta u(t+N-k) \end{bmatrix}^T
\]

\[
\bar{y} = \begin{bmatrix} \bar{y}_1(t) & \cdots & \bar{y}_N(t) \end{bmatrix}^T
\]

The coefficients of \( R_d \) are the first \( d-k+1 \) terms of the impulse response of \( Q^{-k}B/(A\Delta) \), which are the same as the first \( d-k+1 \) terms of the step response of \( q^{-k}B/A \). The matrix \( R \) is thus a lower triangular matrix:

\[
R = \begin{bmatrix} r_0 & 0 & \cdots & 0 \\
r_1 & r_0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
r_{N-1} & r_{N-2} & \cdots & r_0
\end{bmatrix}
\]

If there is a time delay in the system \( (k > 0) \) then the first \( k \) rows of \( R \) will be zero.

Also introduce the vector of desired future outputs:

\[
y_m = \begin{bmatrix} y_m(t+1) & \cdots & y_m(t+N) \end{bmatrix}^T
\]

The expected value of the loss function can be written as:

\[
J = E \left\{ (y - y_m)^T(y - y_m) + \rho \Delta u^T \Delta u \right\}
\]

\[
= (R \Delta u + \bar{y} - y_m)^T(R \Delta u + \bar{y} - y_m) + \rho \Delta u^T \Delta u
\]

Minimization of this expression with respect to \( \Delta u \) gives:

\[
\Delta u = (R^T R + \rho I)^{-1} R^T (y_m - \bar{y})
\]

\[(5.95)\]
The first component in $\Delta u$ is $\Delta u(t)$, which is the control signal applied to the system. The controller automatically has an integrator.

To decrease the computations, it is possible to introduce constraints on the future control signal. For instance, it can be assumed that the control increments are zero after $N_u < N$ steps, where $N$ is the prediction horizon in the loss function:

$$\Delta u(t + d - 1) = 0, \quad d > N_u$$

This implies that the control signal is assumed to be constant after $N_u$ steps. The control law then changes to:

$$\Delta u = (R_1^T R_1 + \rho I)^{-1} R_1^T (y_m - \bar{y})$$  \hspace{1cm} (5.96)

where $R_1$ is the $N \times N_u$ matrix:

$$R_1 = \begin{pmatrix}
  r_0 & 0 & \cdots & 0 \\
  r_1 & r_0 & \cdots & 0 \\
  \vdots & \ddots & \ddots & \vdots \\
  \vdots & & \ddots & r_0 \\
  r_{N-1} & r_{N-2} & \cdots & r_{N-N_u}
\end{pmatrix}$$

The matrix to be inverted is now of the order $N_u \times N_u$. The output and control horizons can be chosen as follows:

$N_1$: If the time delay is known, then $N_1 = k + 1$; otherwise choose $N_1 = 1$.

$N_2$: The maximum output horizon $N_2$ can be chosen such as $N_2 T_s$ is of the same magnitude as the rise time of the system.

$N_u$: Should be at least equal to the number of unstable or poorly damped poles.

When the delay is estimated, only the output predictions for $d > \hat{k} + 1$ are computed and also the first $\hat{k}$ rows in $R_1$ are set equal to zero and this significantly reduces the computations of the control design.
Chapter 5. VRE-Based Adaptive Control

The response of the adaptive GPC/FMVRE is shown in Figure 5.30, for the following parameters:

\[ N_1 = \hat{k} + 1 \]
\[ N_2 = \hat{k} + 4 \]
\[ N_u = 1 \]
\[ \rho = 0.001 \]
\[ y_m(t + d) = \text{present reference output} \]

It is noted that the delay estimation converges faster than for the adaptive PP/FMVRE and the adaptive Dahlin/FMVRE which means a shorter learning period when the system changes. The model parameter estimates are shown in Figure 5.31. The response of the adaptive GPC based on the extended numerator polynomial is shown in Figure 5.32 for \( N_1 = 1 \) and \( N_2 = 12 \). Unlike the PP and the Dahlin controllers, the extended numerator polynomial method will not give the same good response as the VRE-based GPC design after the parameters converge. The reason for this difference is that the design of the GPC depends on the choice of \( N_1 \) and \( N_2 \). In the VRE-based design, \( N_1 \) and \( N_2 \) change as the estimated delay changes, which means that, after the delay, we get essentially the same closed-loop response for the different values of system delay. On the other hand, if the extended numerator polynomial method is used, \( N_1 \) has to be fixed at a value corresponding to the minimum expected system delay and \( N_2 \) has to be fixed at a value corresponding to the maximum expected system delay. This means that the closed-loop response gets slower as the system delay becomes smaller. If the system delay is much less than the maximum expected delay, the adaptive GPC based on the extended numerator method will give the system open-loop response. The adaptive GPC/VRE has the advantages of shorter learning period and consistency of the good closed-loop response for different system delays. The adaptive GPC/EMVRE
has as good a closed-loop response as the adaptive GPC/FMVRE as shown in Figure 5.33

5.4 Delay Estimation in Closed-Loop

5.4.1 Deterministic Systems

When the parameter estimation is carried out in closed-loop, the problem of parameter identifiability arises (Ljung et al. 1974, Söderström et al. 1975, 76 and Gustavsson et al. 1977). Let the system be described by the regression model:

\[ \dot{y}(t) = \Phi(t)\Theta \quad (5.97) \]

Since we have measurements from \( t = 1 \) to the present time, the following vector equation results:

\[ Y(t) = \Phi(t)\Theta \quad (5.98) \]

where:

\[ Y(t) = [y(t) \ y(t-1) \ \cdots \ y(1)]^T \quad (5.99) \]

\[ \Phi(t) = \begin{bmatrix} \Phi(t) \\ \Phi(t-1) \\ \vdots \\ \Phi(1) \end{bmatrix} \quad (5.100) \]

and the Least-Squares parameter estimate is given

\[ \hat{\Theta} = (\Phi^T\Phi)^{-1}\Phi^TY \quad (5.101) \]

The matrix \((\Phi^T\Phi)\) is called the excitation matrix and it should be invertible which means that the columns of \(\Phi(t)\) should be linearly independent. In open-loop estimation this means that the system input should be persistently exciting. In closed-loop
Figure 5.30: Adaptive GPC/FMVRE response and delay estimation
Figure 5.31: Adaptive GPC/FMVRE parameter estimates
Figure 5.32: Adaptive GPC/extended polynomial response
Figure 5.33: Adaptive GPC/EMVRE response and delay estimation
estimation the persistent excitation is required, but also $u(t)$ should not be linearly
dependent only on the other parameters in $\Phi(t)$ through the feedback path. This im-
poses constraints on the controller structure and if these constraints are violated a
non-unique parameter estimate results.

The delay estimation using VRE does not include the inverse of the excitation
matrix. The condition that the excitation matrix should be invertible can be violated
and still we get a unique value of delay estimation. However, as was seen in the previous
chapter, a certain excitation level is required for an unbiased delay estimation. The
excitation required for an unbiased delay estimation is not the same as that required
for unbiased parameter estimation. An excitation level as low as one sinusoid can be
enough for unbiased delay estimation for any system order as long as the system pole
excess is one, while one sinusoid is not persistently exciting if the system has more than
two parameters. The closed-loop delay estimation in the previous section shows that
the delay does not lose identifiability as we close the loop and the estimated delay is
exactly the same as that obtained by open-loop estimation.

5.4.2 Stochastic Systems

If process noise exists, the output can be expressed as in Equation 4.49:

$$y(t) = G(q^{-1}) u(t) + G_e(q^{-1}) e(t)$$

$$y_u + y_e$$

and the input-output cross-correlation is giving by:

$$r_{uy}(t) = r_{uy_u}(t) + r_{uy_e}(t)$$

The term $r_{uy_e}$ represents the cross-correlation between the present input and the future
process noise. In open-loop estimation, the input is not correlated with the process
noise and hence we get $r_{uy_e} = 0$, which means that the process noise has no effect on the delay estimation.

In closed-loop estimation, the input is indeed correlated with the past process noise through the feedback path. For input given by:

$$R(q^{-1})u(t) = T(q^{-1})u_c(t) - S(q^{-1})y(t)$$

the system input in closed loop is given by:

$$u = \frac{GS/T}{1 + GS/R} u_c + \frac{GS/R}{1 + GS/R} y_e$$  \hspace{1cm} (5.102)

$u_c$ is not correlated with $y_e$ and hence it does not contribute to the term $r_{uy_e}$. The present input is related to the past $y_e$ by the closed-loop transfer-function. This means that the present $y_e$ is related to the future $u$ by the closed-loop transfer-function, the term $r_{y_eu}$ can then be written as:

$$r_{y_eu}(t) = \frac{GS/R}{1 + GS/R} r_{y_e}(t)$$  \hspace{1cm} (5.103)

The function $r_{y_eu}$ is generated from the auto-correlation of the process noise $r_{y_e}$ through the closed-loop transfer function and then the function $r_{uy_e}$ can be obtained as a mirror image of $r_{y_eu}$ around the zero shift axis ($r_{uy_e}(t) = r_{y_eu}(-t)$), this is explained in Figure 5.34. The function $r_{uy_e}(t)$ does not, in general, have a zero value for $t \geq 2k_{\text{min}}$, which means that we may, or may not, get a bias in the delay estimation because of the non-zero value of $r_{uy_e}(t)$.

If the process noise $y_e$ has a non-zero auto-correlation for only $t \leq 2k_{\text{min}}$, then $r_{uy_e}(t) = 0$, $t \in [k_{\text{min}}, k_{\text{max}}]$ and no bias in the delay estimation results. A white process noise is one special case where $y_e$ has a non-zero auto-correlation only for $t = 0$ and we conclude that such a process noise does not affect the delay estimation in closed-loop. The closed-loop system does not have to be stable to apply this conclusion. If the closed-loop is unstable, $r_{y_eu}(t)$ will grow indefinitely, but not $r_{uy_e}(t)$.
Figure 5.34: Correlation functions of the process noise
If the process noise does not satisfy the above condition on its auto-correlation, a delay estimation bias may result and one of the following options may be taken to eliminate this problem:

- If the process noise has known properties in the bandwidth of interest, the input and output signals can be filtered through the inverse of the estimated noise filter before the estimation is carried out. This approach is similar to that taken by Clarke et al. (1987-a,b) in the design of the GPC. In this case:

\[
\begin{align*}
(G^y(t)) &= G(G^{-1} u(t)) + G^{-1} G e(t) \\
y_f(t) &= G u_f(t) + \tilde{y}_e
\end{align*}
\]  

(5.104)

\(y_f\) and \(u_f\) are the signals used in delay estimation. The filtered process noise \(\tilde{y}_e\) should now satisfy the required condition on its auto-correlation.

- If the process noise has unknown properties we can replace the input signal \(u(t)\) used in the delay estimation by the command signal \(u_c(t)\) and an unbiased delay estimation should result. This is based on the fact that the delay of the open-loop system equals the delay of the closed-loop system (no delay is arbitrary inserted in the controller). The control variable \(u_c\) is uncorrelated with \(y_e\) and hence the process noise has no effect on the delay estimation. This method is equivalent to operating the VRE in an auto-tuning mode instead of the self-tuning mode.

- EMVRE can be used with the Approximate Maximum Likelihood to estimate the system parameters and the noise filter parameters.

5.5 Conclusion

In this chapter we showed how to utilize the direct delay estimation obtained from the VRE to design controllers with the minimum number of parameters. The closed-loop
response is much better than that obtained without the VRE. The design of VRE-based Pole Placement, Dahlin and Generalized Predictive Controllers is given. The direct delay estimation from the VRE allowed us tune the GPC and get the same closed-loop performance for any system delay, this was not possible otherwise.

Unlike the system parameters, the delay cannot lose identifiability in closed-loop identification. For stochastic systems, a bias in the delay estimation may result. This bias can be avoided by filtering the input and output signals through the inverse of the noise filter, if known. If the noise filter is not known, the delay estimation can be carried out in the auto-tuning mode or by estimating the noise filter parameters using the Approximate Maximum Likelihood.
Chapter 6

Experiments on Delay Estimation

In order to demonstrate the efficiency of the new delay estimation method, two experiments are carried out. In the first experiment, the FMVRE is used to estimate the model of a temperature control experiment. The estimated model is used in the adaptive Dahlin controller design. The experiment demonstrates the efficiency of the FMVRE in closed-loop adaptive control. In the second experiment, the EMVRE is used to estimate the delay of a pulp refiner industrial process. The experiment demonstrates the efficiency and speed of the EMVRE with noisy measurements.

6.1 Temperature Control Experiment

The schematic diagram of the experimental setup is shown in Figure 6.35. A similar setup was built by Gendron et al. 1990-a, b. The setup consists of a tank where cold water is heated and then passed through a series of three long hoses. The water temperature at each hose end is measured and can be used as the process output. The process input is the electrical energy delivered to the 110V, 1500W heater through the electronic power switch (Triac). The process is controlled by an IBM-PC compatible computer and the software is written in Microsoft Quick Basic.

The cold water comes from the domestic water line. As expected, the temperature and the flow rate of this water change significantly with time. The change in the cold water temperature is a disturbance to the process. Flow rate changes directly change the process dynamics and delay. The hoses used are simple 1/2", 50' garden hoses.
Figure 6.35: Temperature control experiment
There is a significant heat loss from the hoses due to convection. This is another disturbance to the system. The sampling period used is 42.5 sec. and this results in a time delay of about four samples in every hose. The maximum value for the delay search was set to 15 samples.

The gating circuit turns-on the Triac for an integer number of the power supply cycles and the duty cycle of the Triac is controlled from 0% to 100%. The computer controls the duty cycle through an 8-bit output port. The temperature is measured by the LM335A solid state devices which provide linear temperature measurements to the computer through the A/D ports.

The model used is a second-order model of the form:

\[ y(t + 1) = -\theta_1 y(t) - \theta_2 y(t - 1) + \theta_3 u(t - \hat{k}) + \theta_4 u(t - 1 - \hat{k}) + \theta_5 \]

The bias term \( \theta_5 \) is used to represent the operating point changes due to the cold water temperature changes and the convection heat loss through the hoses. The parameters \( \theta_1 \) to \( \theta_5 \) are estimated using RLS and \( \hat{k} \) is estimated using FMVRE.

Figure 6.36 shows the experimental results over a period of 1600 samples i.e. about 19 hours. The controlled output is switched to a different temperature sensor about every 400 samples. The controlled output is required to track a square wave variation in the reference signal. Since we are mainly interested in delay estimation results, the adaptive Dahlin controller is used. A more complicated controller will result in a better closed-loop response using the same estimation model.

The results show excellent convergence and speed for the delay estimation. As the estimated delay gets very large, it can oscillate between two adjacent values instead of converging to one of them. The reason is that the function \( E_1 \) becomes flatter near its maximum if the excitation signal is not rich enough or if the forgetting factor is small. However, both delay values represent the process reasonably well, keeping in mind that
Chapter 6. Experiments on Delay Estimation

Figure 6.36: Delay estimation and temperature variation
any alternative method would require many more parameters to represent the process for such a large value of the delay. The controlled temperature shows some offset in the first 400 samples in Figure 6.36. The Dahlin controller has an integrator which should be able of eliminating this offset. The period of the reference square wave signal is not long enough to let the system reach its steady-state value. After the first 400 samples, the estimator has a better model of the system and it eliminates the offset more efficiently. The input signal shows significant changes in its average values over the experiment duration. This is required to counteract the changes in the heat loss by convection and the changes in the ambient temperature. It is important to note that the output average value does not change over the experiment duration.

6.2 Reject Pulp Refiner Experiment

The use of chip and pulp refiners is spreading in the wood pulp industry, especially in Thermo-Mechanical Pulping (TMP) (Dumont et al. 1982-c and 1988). A reject pulp refiner consists of two counter-rotating grooved plates, with pressure exerted on one of them by a hydraulic cylinder as in Figure 6.37 (Dumont 1982-a). Wood pulp and dilution water are fed near the axis and forced to move outward between the plates by centrifugal and frictional forces. Steam and pulp are discharged at the periphery. The specific energy, or energy per mass unit of wood fibers, is a major factor controlling the pulp quality. This means that the feed-rate and the motor load must be controlled. To control the motor load, the plate gap is manipulated by adjusting the hydraulic pressure.

The process input is the gap between the two plates and the process output is the motor load in $MW$. A generated PRBS is added to the gap reference setting. The sampling time is $5\text{sec.}$ and the duration of the pulse is sent to a micro-dial which sends
Figure 6.37: Schematic representation of a pulp refiner.
Chapter 6. Experiments on Delay Estimation

a signal to a servo valve directing oil to the cylinder.

The input and output measurements are shown in Figure 6.38 and the model delay and gain estimation are shown in Figure 6.39. The delay estimation converges to the value of one in less than 20 samples, which is extremely fast for this noisy measurements. The process model is known a-priori to be pure gain plus delay and the delay is in the range of $k \in [0,5]$. Because the exact process gain and delay are not known, the EMVRE delay estimation is compared with that obtained from off-line analysis of the same measurements. The input auto-correlation and the input-output cross-correlation functions are shown in Figure 6.40. There is a shift of two samples between the first non-zero auto-correlation value and the first non-zero cross-correlation value, assuming that the threshold of zero and non-zero is 10% of the maximum value obtained. This means a delay of one sample, i.e.

$$y(t + 1) = k_y u(t - 1).$$

and this value agrees with the EMVRE estimation.

6.3 Conclusions

In this chapter we presented experimental results for closed-loop and open-loop delay estimation. Adaptive Dahlin/FMVRE is used to control the temperature in an experiment having long delay. The results show fast adaptation and good closed-loop performance over the wide range of delay variation. In the second experiment EMVRE is used to model the Reject Pulp Refiner. Although the measurements are very noisy, the delay estimation converges in less than 20 samples to the same value obtained by off-line analysis. The experiments demonstrate the applicability of the VRE delay estimation methods.
Chapter 6. Experiments on Delay Estimation

Figure 6.38: Measurements on the pulp refiner
Chapter 6. Experiments on Delay Estimation

Figure 6.39: Delay and gain estimation of the pulp refiner
Chapter 6. Experiments on Delay Estimation

Figure 6.40: Correlation functions of the pulp refiner measurements
Chapter 7

Conclusions

7.1 Summary of the Work

In this thesis we have discussed new methods for direct estimation of the unknown delay of a system utilizing as much as possible the information available a-priori. We developed the Variable Regression Estimation which estimates the system delay interactively with the other process parameters (the Estimated Model Variable Regression Estimation) or which estimates the system delay independently of the process parameters (the Fixed Model Variable Regression Estimation). The theoretical study, computer simulations and experimental results show the superior properties of the proposed methods over the traditional delay estimation methods. The results of the thesis are summarized as follows:

- The Variable Regression is introduced and joined with standard parameter estimators. The parameter vector has exactly the same number of unknowns as in the known delay case. No additional parameters are used to represent the delay.

- The Estimated Model Variable Regression Estimation is introduced for systems of small range of parameter variations. The method has only one convergence point corresponding to the correct delay. Coloured noise does not affect the delay estimation.
• The Fixed Model Variable Regression Estimation is introduced for systems of unknown parameters. The estimated delay is accurate for first and second-order systems. The estimated delay approximates the high order dynamics for higher-order systems and the estimated delay is close to that obtained by the Ziegler-Nichols approximation. The FMVRE can also be applied to systems with fractional delay.

• Proofs of the FMVRE properties are derived. The estimated delay is consistent, bounded and unbiased even in the presence of coloured noise. The estimated model is stable and matches the system frequency response at a given frequency in the presence of unmodelled dynamics. The excitation richness required for correct delay estimation is less than the excitation richness required to estimate the other process parameters. One sinusoid can be enough for correct delay estimation. For known parameters system, any non-steady excitation signal is enough for correct delay estimation.

• New adaptive versions of the pole placement, Dahlin and generalized predictive controllers are presented. The controller designs are based on the VRE which results in minimum number of the controller parameters, faster adaptation and good closed-loop response. The adaptive GPC/VRE has the same performance for different system delays. This is not the case without direct delay estimation.

• Unlike the other model parameters, the delay cannot loose identifiability in closed-loop. For stochastic systems, a bias may occur in closed-loop delay estimation depending on the noise and the system structure. This can be eliminated by filtering the input and output signals or estimating the delay between the output and the reference signal.
• Experimental results for closed-loop and open-loop delay estimation are presented. Adaptive Dahlin/FMVRE is used to control the temperature in an experiment having long delay. The results show fast adaptation and good closed-loop performance over the wide range of delay variation. In the second experiment EMVRE is used to model the Reject Pulp Refiner. Although the measurements are very noisy, the delay estimation converges in less than 20 samples to the same value obtained by off-line analysis. The experiments demonstrate the applicability of the VRE delay estimation methods.

7.2 Future Research Directions

Further research can be directed to investigate the properties of the EMVRE. This should study the delay estimation equations and the parameter estimation equations simultaneously. The EMVRE behavior with high-order systems is not fully understood yet. Further research can also be directed to study the delay estimation in closed-loop and derive the condition on the noise required to guarantee unbiased delay estimation. Having done this, the VRE can be used in the design of controllers for stochastic systems such as minimum-variance controllers.

We believe that the VRE-based adaptive controllers can provide excellent control for linear deterministic systems. Applying the developed techniques of estimation and adaptive control to industrial processes having unknown or varying delay is our next target.
Bibliography


