LEAST SQUARES AND INSTRUMENTAL VARIABLE METHODS

Tomas McKelvey
Chalmers University of Technology, SE-412 96 Göteborg, Sweden

Keywords: estimation, system identification, time domain, QR-factorization, singular value decomposition, prediction, discrete time systems, time series analysis, dynamical systems

Contents

Summary

System identification deals with determining models of dynamical systems from measured input and output signals. Model structures in the form of prediction models are particularly suitable both from a conceptual and a statistical point of view. This paper describes two basic estimation methods which calculate estimates of the prediction model parameters from measurements of input and output signals. The simple, yet widely used linear least squares (LS) method and the closely related instrumental variable (IV) method are covered. Both methods are based on a prediction model structure which is *linearly parameterized*.

In the LS method the parameter estimate minimizes the sum of the squared prediction errors between the prediction model and the output data. The IV method is a slight variation of the LS method which has the ability to counteract the effects of a more general class of noise signals which if using the LS method would make the parameter estimate biased. The statistical background regarding bias and variance is covered in depth for the simple LS case when the predictor is independent of the noise. For the more general case when the predictor involves past output measurements we only

comment on results from the literature. Conceptually they resemble the simple case. Details are provided on how to numerically calculate estimates and to provide insight into detecting and handling numerically ill-conditioned situations.

1. Introduction

The least squares method (LS) together with the closely related instrumental variable method (IV) are the most common methods for solving a wide range of estimation problems occurring in numerous applications. Examples can be found in astronomy, physics, economics, medicine, and of course, in most engineering applications. A standard usage is that a particular phenomenon is to be modeled by means of a mathematical model. A parameterized model structure is selected defining a feasible model set. Based on empirical knowledge of the phenomenon under study, often in the form of measurements, a model is selected from the model set based on a model selection criterion.

If the model selection criterion is a minimization of a quadratic function the LS and IV methods can often be applied. Both methods share the important property that the parametric estimate of the model parameters can be found by numerically stable computations with predictable numerical errors and manageable computational complexity. The focus of this chapter is the estimation of parameters of dynamical systems which is an important task in the model based control engineering. A linear regression framework will be adopted where predictor models are linear functions of the unknown parameters.

Digital processing of signals implies that the data is always sampled. Hence, the focus here will be on estimation of discrete time linear systems from sampled data, see also (*Discrete-Time, Sampled-Data, Digital Control Systems, Quantization Effects*) and (*Discrete-Time Equivalents to Continuous-Time Systems*). If the underlying application dictates a continuous time model several routes are possible. The LS and IV techniques presented here can also be reformulated to directly estimate continuous time or delta operator models by prefiltering the data and forming regressors in a different fashion. Please refer to (*Frequency Domain System Identification*) and (*Continuous-time Identification*) for further details.

The estimation of the model parameters can be done in two different ways. In batch-type estimation all available data in the data set is used to determine an estimate. For on-line applications where a model continuously needs to be updated, a recursive methodology is then very common. When a new data sample is available the estimate of the model parameters is updated based on the new data and the past estimate and data. Both the LS and the IV methods can be implemented as batch-type algorithms or as recursive algorithms. In this chapter only the batch-type will be presented and the reader is referred to (*Recursive Algorithms*) for a deeper discussion on the recursive algorithms.

The history of estimation and system identification dates back to the fundamental work by Gauss who introduced the least squares method for fitting a model to measured data corrupted by noise. The theory of estimation has continuously grown during the 20th

century in many disciplines. A full understanding of the estimator properties for the case when the predictor contains time delayed versions of both inputs and noisy outputs was clarified within the control community during the active era approximately during 1965-1980.

2. Models as Predictors

The notion of a model is closely related to prediction, i.e., given certain knowledge predict the outcome of a process. In the area of automatic control a model is a mathematical operator which maps the space of input signals to the space of output signals. With the notation $u(t)$ as the input signal, G as the operator representing the system and $y(t)$ as the output we write

$$y(t) = Gu(t) \tag{1}$$

This chapter will entirely focus on the special case when the operator G is a time invariant discrete time linear system. The relation (1) can then be expressed more explicitly. By introducing the impulse response sequence of the system as $g(k),\ k = 1,\ldots,\infty$ we obtain

$$y(t) = \sum_{k=1}^{\infty} g(k)u(t-k) \triangleq \sum_{k=1}^{\infty} g(k)q^{-k}u(t) \triangleq G(q)u(t) \tag{2}$$

The output at time t is a linear combination of all past inputs. In (2) we also have introduced the time shift operator q^{-1}. The object $G(q)$ is called the transfer function of the system model. For further discussion on transfer functions refer to (*General Models of Dynamical Systems* and *Description of Continuous Linear Time-Invariant Systems in Time-Domain*). The model (2) lacks a representation of disturbances and/or unmeasured inputs acting on the system and therefore also present in the output. A more complete model is illustrated in Figure 1 where a noise term $w(t)$ is added.

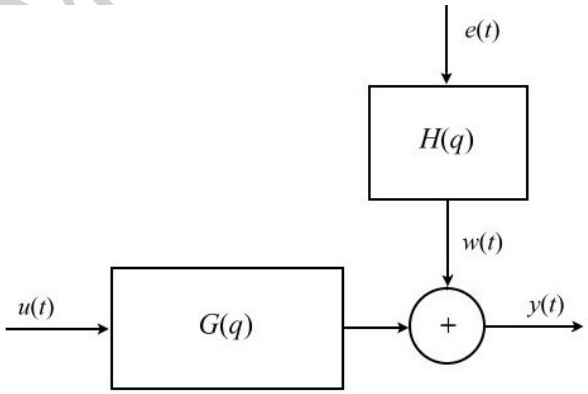

Figure 1: System setup. The input $u(t)$ is assumed known and the noisy output $y(t)$ is measured and is the sum of the deterministic part $G(q)u(t)$ and the stochastic part $w(t) = H(q)e(t)$ where $e(t)$ is unknown but assumed to have zero mean..

The noise is modeled as $w(t) = H(q)e(t)$ where $H(q)$ is a linear transfer function and $e(t)$ is taken as a sequence of independent identically distributed (i.i.d.) random variables of zero mean and variance λ. The transfer function $H(q)$ is used in order to include descriptions where the noise $w(t)$ is correlated (or colored) in time. We further assume that $H(q)$ is stable, inversely stable and monic. (Write $H(q) = h_0 + h_1 q^{-1} + \ldots$. If $h_0 = 1$ then $H(q)$ is monic.) A transfer function $H(q) = \sum_{k=0}^{\infty} h_k q^{-k}$ is stable if the impulse response $h(k)$ is absolutely summable, that is $\sum_{k=0}^{\infty} |h_k| < \infty$. Putting the deterministic and stochastic parts together yields

$$y(t) = G(q)u(t) + w(t) = G(q)u(t) + H(q)e(t) \tag{3}$$

See Chapters (*Control of Stochastic Systems* and Models of Stochastic Systems) for more related material on models of stochastic systems.

Throughout the chapter we assume that $G(q)$ is stable and the input $u(t)$ is *independent* of the output $y(t)$ and is *not* generated by output feedback by a closed loop control system. Identification under feedback conditions normally requires special care. Methods for the closed loop case can be found in (*Prediction Error Method*, *Identification for Control* and *Frequency Domain System Identification*).

A simple reorganization of (3) to

$$y(t) = H(q)^{-1}[G(q)u(t) + (H(q) - 1)y(t)] + e(t) \tag{4}$$

reveals the structure of the one-step ahead predictor. Note that $(H(q) - 1)y(t) = \sum_{k=1}^{\infty} h_k y(t-k)$ (due to the monicity of $H(q)$) is only dependent on past outputs. As $e(t)$ is unknown at time $t-1$, zero mean and independent, a natural predictor for $y(t)$ is obtained by simply assuming $e(t) = 0$

$$\hat{y}(t) \stackrel{\Delta}{=} H(q)^{-1}\left[G(q)u(t) + \left(H(q) - 1\right)y(t)\right] \tag{5}$$

Given a predictor the *prediction errors* are defined as $\varepsilon(t) = y(t) - \hat{y}(t)$. It can be shown that the prediction error produced by the predictor $\hat{y}(t)$ in (5) indeed has the lowest variance for all linear one-step ahead predictors for $y(t)$ defined in (3). For further information about prediction models please refer to (*Identification of Linear Systems in Time Domain* and *Prediction Error Method*).

2.1. Linearly Parameterized Predictors

The aim of the system identification is to determine the unknown transfer function $G(q)$ and perhaps also $H(q)$ from samples of the input and output signals. This task consists in two steps. First a model set is defined, e.g. by deciding on the model

order(s). In the second step the model parameters are determined by an estimation procedure. This chapter focuses on the second step, the estimation of the parameters. In order to accomplish this task we design a parameterized model structure of the predictor (5) and look for parameters which minimize the prediction errors. For some particular choices of model structures, this task is considerably simplified and the parameters can be solved for by analytical methods, e.g. the least squares and the instrumental variable methods.

A range of different model structures emerges when describing the system transfer function $G(q)$ and noise transfer function $H(q)$ in various ways. An assumption that the impulse response of the system has a finite duration is for example very common in models of communication channels. This implies that $G(q) = B(q) = b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b}$ and if we also assume the noise transfer function is unity, i.e., $H(q) = 1$, we obtain the simple finite impulse response (FIR) model structure

$$y(t) = B(q)u(t) + e(t) \tag{6}$$

The predictor (5) associated with the FIR model is trivially given by

$$\hat{y}(t) = B(q)u(t) \tag{7}$$

By introducing a vector of parameters

$$\theta^T = [b_1, b_2, \ldots, b_{n_b}]$$

and a regression vector $\phi(t)$

$$\phi(t)^T = \left[u(t-1), u(t-2), \ldots, u(t-n_b)\right] \tag{8}$$

the *parameterized predictor* can be written as a linear regression

$$\hat{y}(t|\theta) = \phi(t)^T \theta \tag{9}$$

where the dependence on the parameters is made explicit in the notation of the predictor.

Resonant mechanical structures have an infinite impulse response and are difficult to model using moderately sized FIR models. Transfer functions with poles (A transfer function $H(z)$ has a pole at z_p if $H(z)$ has a singularity at $z = z_p$) will be better to approximate such systems. Also if the structure is excited by unmeasured inputs regarded as noise, the noise model should also have a similar pole structure. If we let both the system and the noise transfer function share the same pole polynomials we obtain the ARX structure

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \quad \text{with} \quad \begin{aligned} A(q) &= 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a} \\ B(q) &= b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b} \end{aligned} \tag{10}$$

For a mechanical structure the poles, i.e. the roots of $A(q)$, will contain the resonance frequencies and damping coefficients of the oscillatory modes of the structure. The name ARX stands for Auto Regressive with eXternal input.

One prime advantage with the ARX structure becomes apparent if we consider the predictor (5) using the ARX structure of the models. This leads to the predictor

$$\hat{y}(t|\theta) = -\sum_{k=1}^{n_a} a_k y(t-k) + \sum_{k=1}^{n_b} b_k u(t-k) = \left[1 - A(q)\right] y(t) + B(q)u(t) = \phi(t)^T \theta \tag{11}$$

where

$$\begin{aligned} \phi(t)^T &= \left[u(t-1),\ u(t-2),\ldots,\ u(t-n_b),\ -y(t-1),\ -y(t-2),\ldots,\ -y(t-n_a)\right] \\ \theta^T &= \left[b_1, b_2, \ldots, b_{n_b}, a_1, a_2, \ldots, a_{n_a}\right] \end{aligned} \tag{12}$$

The predictor (11) is a *linear* function of the n_a past outputs and the n_b past inputs and is also a *linear* function of the polynomial coefficients of the system. Hence, the predictor for an ARX model is a function of finite amount of past data.

If only the system dynamics is of importance, it is then possible to neglect the exact form of the noise dynamics. It is then convenient to assume an ARX model of the form

$$A(q)y(t) = B(q)u(t) + v(t) \tag{13}$$

where the ARX equation error is given by $v(t) = A(q)w(t) = A(q)H(q)e(t)$. If the simple ARX predictor given in (11) is used, the predictor clearly is not variance optimal since the true noise dynamics have been neglected. We will later also see that a special estimation technique must be employed to obtain consistent parameter estimates since the equation error $v(t)$ in (13) is not white.

A special case emerges when the system has no input. The output is then equal to the noise, $y(t) = H(q)e(t)$. We will consider here only the autoregressive (AR) case with noise model $H(q) = 1/A(q)$ and prediction formula

$$\hat{y}(t|\theta) = -\sum_{k=1}^{n_a} a_k y(t-k) = \left[1 - A(q)\right] y(t) \tag{14}$$

Note that the AR case is simply the ARX case without the exogenous input $u(t)$.

The FIR, ARX and AR model structures share an important common feature. The one-

step ahead predicted output can be written as a linear regression

$$\hat{y}(t|\theta) = \phi(t)^T \theta \tag{15}$$

where the regression vector $\phi(t)$ contains only a finite amount of past measured data up to time $t-1$.

3. Estimating the Model Parameters

System identification aims at finding a model based on samples of the input and output signals. An immediate question is how to determine the model parameters from data? This chapter deals with two particular techniques; the least squares method and the instrumental variable method. The motivations behind these can be drawn from a statistical point of view where noises and signals are assumed to be quasi-stationary.

Assume we want to find a way to predict the output $y(t)$ of system given knowledge of a set of variables, the regressors $\phi(t)$. If we constrain the prediction function to be linear function parameterized by a vector θ, it can be described as

$$\hat{y}(t|\theta) = \phi(t)^T \theta$$

A rather natural way of finding out the best parameter vector θ would be to minimize the size of the prediction error $y(t) - \hat{y}(t|\theta)$ with respect to the parameters. A possible choice of characterizing this size is the variance of the prediction error. This leads to the minimum variance estimate

$$\hat{\theta}_{MV} = \arg\min_{\theta} \mathrm{E}\left\{[y(t) - \phi(t)^{\mathrm{T}}\theta]^2\right\} \tag{16}$$

where $\mathrm{E}\{\cdot\}$ denotes expectation. However, in reality we only have access a finite set of data and the minimum variance criterion cannot be evaluated. Given N pairs of data, $\{u(t), y(t)\}_{t=1}^{N}$ of data we instead form the sample estimate of the variance as

$$V(\theta) \stackrel{\Delta}{=} \frac{1}{N}\sum_{t=1}^{N}\left\|y(t) - \hat{y}(t,\theta)\right\|^2 = \frac{1}{N}\sum_{t=1}^{N}\left\|y(t) - \phi(t)^T\theta\right\|^2 \tag{17}$$

and define the least square estimate (LSE) as

$$\hat{\theta}_{LS} \stackrel{\Delta}{=} \arg\min_{\theta} V(\theta) \tag{18}$$

Does the LSE $\hat{\theta}_{LS}$ make sense? Clearly if we are looking for a linear function which can predict the output, the resulting predictor $\hat{y}(t|\hat{\theta}_{LS}) = \phi(t)^T\hat{\theta}_{LS}$ is the one which, for the given data and regression vector, does the best job in a least squares sense. This is of

course true even if the predictor structure does not match how the original system generated the data. If the resulting predictor also will perform well on new data not used to determine $\hat{\theta}_{LS}$ depends solely on the validity of the assumed model structure and on the statistical properties of the noise signal $v(t)$. Under mild assumptions on the data the LS estimate will converge to the minimum variance solution (16) as the number of samples tends to infinity. We will also examine the conditions which ensure that as $N \to \infty$ we obtain $\hat{\theta}_{LS} \to \theta_0$ where θ_0 is the true parameters of the system.

A second motivation for the LS estimate is its close relation to the maximum-likelihood estimate. Assume the data generating system can be described by

$$y(t) = \phi(t)^T \theta_0 + e(t)$$

where $e(t)$ is a sequence of i.i.d. random variables with zero mean drawn from a normal distribution. Minimizing the LS-criterion $V(\theta)$ in (17) coincides with maximizing the parameterized likelihood function for the observed data. In Section 4 we will further analyze the LS estimate in some detail.

-
-
-

Bibliography

Chung K.L. (1974). *A Course in Probability Theory*. Academic Press, San Diego, CA. [This book contains many useful results on convergence of sums of random variables].

Draper N.R., Smith H. (1981). *Applied Regression Analysis.* New York: Wiley, 2nd edn. [Good reference on classical regression analysis].

Gauss K.F. (1963). *Theoria Motus Corporum Celestium*: *English Translation: Theory of the Motion of Heavenly Bodies.* New York: Dover. [The root of the least-squares parameter estimation technique].

Golub G.H., Van Loan C.F. (1989). *Matrix computations*, Baltimore, Maryland: The Johns Hopkins University Press, 2nd edn. [This text gives a good account on the numerical issues of the least-squares problem and presents numerically efficient algorithms].

Hannan E.J., Deistler M. (1988). *The Statistical Theory of Linear Systems.* Wiley. [This book gives a thorough theoretical presentation of system identification for linear systems].

Ljung L. (1999). *System Identification: Theory for the User.* Englewood Cliffs, New Jersey: Prentice-Hall, 2nd edn. [This is an advanced textbook on the graduate level which covers most aspects of system identification. A proof of Theorem 1 is given in Chapter 2].

Söderström T., Stoica P. (1983). *Instrumental Variable Methods for System Identification*. Lecture Notes in Control and Information Sciences, New York: Springer-Verlag. [Contains an extensive analysis of various instrumental variable estimators].

Söderström T., Stoica P. (1989), *System Identification.* Hemel Hempstead, Hertfordshire: Prentice-Hall International. [This textbook gives a comprehensive treatment of the system identification subject].

Wong K.Y., Polak E. (1967). Identification of linear discrete time systems using the instrumental variable approach. *IEEE Transactions on Automatic Control*, **AC-12**, 707-718. [The first consistency analysis of the IV method applied to dynamical systems].

Young P.C. (1976). Some observations on instrumental variable methods of time-series analysis. *International Journal of Control* **23**(5), 593–612. [This article provides a nice historical background of the instrumental variable methods and suggests an iterative refinement technique (Refined IV)].

Young P.C. (1984). *Recursive Estimation and Time Series Analysis.* Berlin: Springer-Verlag. [Gives a detailed treatment of instrumental variable techniques].

Biographical Sketch

Tomas McKelvey was born in 1966 in Lund, Sweden. He received his M.Sc. degree in Electrical Engineering from Lund Institute of Technology, Sweden in 1991 and his Ph.D. degree in Automatic Control from Linköping University, Sweden in 1995. From 1995 to 2000 he has held Assistant and Associate professor positions at Linköping University. Currently he is an Associate professor (docent) of Signal Processing in the department of Signals and Systems at Chalmers Technical University, Göteborg, Sweden. Between 1999 and 2000 he was a visiting researcher at University of Newcastle, Australia. Prof. McKelvey is currently the Chairman of International Federation of Automatic Control's (IFAC) Technical Committee on Modelling, Identification and Signal Processing and is also an associate editor for *Automatica*. His main scientific interests are system identification, spectral estimation, time series analysis and signal processing.