PREDICTION ERROR METHODS

Torsten Söderström
Department of Systems and Control, *Information Technology*, *Uppsala University*, *Uppsala*, *Sweden*

Keywords: prediction error method, optimal prediction, identifiability, ARMAX model, consistency

Contents

Summary

A general linear stochastic model is introduced. It will be described how it covers various typical special cases, like

- ARMAX models and other black box input-output model,
- state space models that can be parameterized either by black box considerations or by physical insight.

The general linear model can be used also to discuss identifiability. *Prediction error methods* are introduced as a general methodology for estimating the parameters in a general linear model. The parameter estimates are obtained by minimization of the

sample prediction error variance. The estimates are consistent and have a low asymptotic covariance matrix under weak conditions.

1. Description

1.1. Introduction

In this subsection, the class of prediction error methods (PEM's) will be described. The description is confined to the off-line (or batch case), where the parameter vector in a general linear model of the following form is estimated:

$$\begin{aligned} M(\theta): \quad & y(t) = G(q^{-1};\theta)u(t) + H(q^{-1};\theta)e(t) \\ & Ee(t)e^T(s) = \Lambda(\theta)\delta_{t,s} \end{aligned} \qquad (1)$$

Here, $y(t)$ is the ny-dimensional output at time t and $u(t)$ the nu-dimensional input. Further, $e(t)$ is a sequence of independent and identically distributed (iid) random variables with zero mean, which is referred to as *white noise*. Further, $G(q^{-1};\theta)$ and $H(q^{-1};\theta)$ are filters of finite order (i.e. rational functions of q^{-1}). They have dimensions $(ny\,|\,nu)$ and $(ny\,|\,ny)$, respectively.

The parameter vector θ is to be estimated from the available input-output data $y(1), u(1), ..., y(N), u(N)$. In practice, one does not use the model Eq. (1) as such, but a special case such as an ARMAX model or some suitably parameterized state space model. To keep the description general, confine it here to the model Eq. (1) which covers deliberately all possible cases.

A model obtained by identification can be used in many ways, depending on the purpose of modeling. In many applications the model is used for *prediction*. Note that this is often inherent when the model is to be used as a basis for control system synthesis. Most systems are stochastic, which means that the output at time t cannot be determined exactly from data available at time $t-1$. It is thus valuable to know, forecast or predict at time $t-1$ what the output of the process is likely to be at time t, in order to take an appropriate control action, i.e. to determine the input $u(t-1)$.

It, therefore, makes sense to determine the model parameter vector θ so that the prediction error

$$\varepsilon(t,\theta) = y(t) - \hat{y}(t\,|\,t-1;\theta) \qquad (2)$$

is small. In Eq. (2), $\hat{y}(t\,|\,t-1;\theta)$ denotes a prediction of $y(t)$ given the data up to and including time $t-1$ (i.e. $y(t-1), u(t-1), y(t-2), u(t-2), ...$) and based on the model parameter vector θ.

Now formalize this idea and consider the general model structure introduced in Eq. (1). Assume that $G(0,\theta)=0$, i.e. that the model has at least one pure delay from input to output. The *optimal mean square predictor* can conceptually be written as

$$\hat{y}(t\,|\,t-1;\theta)=L_1(q^{-1};\theta)y(t)+L_2(q^{-1};\theta)u(t) \tag{3}$$

which is a function of past data only if the predictor filter $L_1(q^{-1};\theta)$ and $L_2(q^{-1},\theta)$ are constrained by

$$L_1(0;\theta)=0, \quad L_2(0;\theta)=0\,. \tag{4}$$

Once the model and the predictor are given, the prediction errors are computed as in Eq. (2). The parameter estimate $\hat{\theta}$ is then chosen to make the prediction errors $\varepsilon(1,\theta),...,\varepsilon(N,\theta)$ small.

To define a prediction error method the user has to make the following choices:

- *Choice of model structure*. This concerns the parameterization of $G(q^{-1};\theta), H(q^{-1};\theta)$ and $\Lambda(\theta)$ in Eq. (1) as functions of θ.
- *Choice of criterion*. This concerns a scalar-valued function of all the prediction errors $\varepsilon(1,\theta),...,\varepsilon(N,\theta)$, which will assess the performance of the predictor used; this criterion is to be minimized with respect to θ to choose the “best” predictor in the class considered.

The most common choice of criterion is (for the single-output case) the sample variance of the prediction errors

$$V_N(\theta)\triangleq\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t,\theta)\,. \tag{5}$$

Several modifications exist, though, see subsection 1.6. For multivariable models $(ny>1)$, the criterion can be modified, for example, as

$$V_N(\theta)=\frac{1}{N}\sum_{t=1}^{N}\|\,\varepsilon(t,\theta)\,\|^2\,, \tag{6}$$

but several other alternatives exist as well.

The prediction error estimate of θ is now defined as the minimizing element of the criterion Eq. (5), i.e.

$$\hat{\theta}=\arg\min_{\theta}V_N(\theta) \tag{7}$$

To complete the description, it remains to describe how the prediction errors are to be computed for the general linear model, Eq. (1). This is done in subsection 1.5.

1.2. General Linear Dynamic Models

1.2.1. Introduction

In this subsection, general linear models of the form Eq. (1) will be discussed. The model will be parameterized by a vector θ, which is to be estimated.

The general form of model structure that will be used is the following

$$\begin{aligned} M(\theta):\ & y(t) = G(q^{-1};\theta)u(t) + H(q^{-1};\theta)e(t), \\ & Ee(t)e^T(s) = \Lambda(\theta)\delta_{t,s}. \end{aligned} \tag{8}$$

The filters $G(q^{-1};\theta)$ and $H(q^{-1};\theta)$ as well as the noise covariance matrix $\Lambda(\theta)$ are functions of the parameter vector θ. Often θ (which is assumed to be $n\theta$ -dimensional) is restricted to lie in a subset of $R^{n\theta}$. This set is given by

$$\begin{aligned} D\ =\ \{\theta \,|\, & H^{-1}(q^{-1};\theta) \text{ and } H^{-1}(q^{-1};\theta)G(q^{-1};\theta) \text{ are asymptotically stable} \\ & G(0;\theta)=0, H(0;\theta)=I, \Lambda(\theta) \text{ is nonnegative definite}\}. \end{aligned} \tag{9}$$

The reasons for these restrictions in the definition of D will become clear in the following, where it will be shown that when θ belongs to D, there is a simple form for the optimal one step prediction of $y(t)$ given past data $y(t-1), u(t-1), y(t-2), u(t-2),....$

For stationary disturbances with rational spectral densities it is a consequence of the spectral factorization theorem, that they can be modeled within the restrictions given by Eq. (9).

Eq. (8) describes a general linear model. The following examples describe typical model structures by specifying the parameterization. That is to say, they specify how $G(q^{-1};\theta), H(q^{-1};\theta)$ and $\Lambda(\theta)$ depend on the parameter vector θ.

1.2.2. ARMAX Models

Let $y(t)$ and $u(t)$ be scalar signals and consider the model structure

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \tag{10}$$

where

$$\begin{aligned} A(q^{-1}) &= 1 + a_1 q^{-1} + \cdots + a_{na} q^{-na} \\ B(q^{-1}) &= b_1 q^{-1} + \cdots + b_{nb} q^{-nb} \\ C(q^{-1}) &= 1 + c_1 q^{-1} + \cdots + c_{nc} q^{-nc}. \end{aligned} \tag{11}$$

The parameter vector is taken as

$$\theta = (a_1 ... a_{na} \quad b_1 ... b_{nb} \quad c_1 ... c_{nc})^T . \tag{12}$$

The model Eq. (10) is called an ARMAX model, which is short for an ARMA model (AutoRegressive Moving Average) with an eXogenous signal (i.e. a control variable $u(t)$ is present). The largest integer of the triple (na, nb, nc) is called the *model order*.

To see how Eq. (10) relates to Eq. (8), note that it can be written as

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})} u(t) + \frac{C(q^{-1})}{A(q^{-1})} e(t) . \tag{13}$$

Thus, for the model structure Eq. (10),

$$\begin{aligned} G(q^{-1};\theta) &= \frac{B(q^{-1})}{A(q^{-1})} \\ H(q^{-1};\theta) &= \frac{C(q^{-1})}{A(q^{-1})} \\ \Lambda(\theta) &= \lambda^2. \end{aligned} \tag{14}$$

The set D is given by

$$D = \{\theta \mid \text{The polynomial } C(z) \text{ has all zeros outside the unit circel}\} . \tag{15}$$

A more standard formulation of the requirement $\theta \in D$ is that the *reciprocal* polynomial

$$C^*(z) = z^{nc} + c_1 z^{nc-1} + \cdots + c_{nc} = z^{nc} C(z^{-1}) \tag{16}$$

has all zeros *inside* the unit circle.

There are several important special cases of Eq. (10):

- An *autoregressive* (AR) model is obtained when $nb = nc = 0$. (Then a pure time series is modeled, i.e. no input signal is assumed to be present.) For this case

$$\begin{aligned} A(q^{-1})y(t) &= e(t) \\ \theta &= (a_1 ... a_{na})^T . \end{aligned} \tag{17}$$

- A *moving average* (MA) model is obtained when $na = nb = 0$. Then

$$\begin{aligned} y(t) &= C(q^{-1})e(t) \\ \theta &= (c_1 ... c_{nc})^T . \end{aligned} \tag{18}$$

- An *autoregressive moving average* (ARMA) model is obtained when $nb = 0$. Then

$$\begin{aligned} A(q^{-1})y(t) &= C(q^{-1})e(t) \\ \theta &= (a_1 ... a_{na} \quad c_1 ... c_{nc})^T . \end{aligned} \tag{19}$$

When $A(q^{-1})$ is constrained to contain a factor $1 - q^{-1}$ the model is called autoregressive integrated moving average (ARIMA). Such models are useful for describing drifting disturbances.

The above three special cases do all apply to a pure time series (there is no input signal). Some other special cases are:

- A finite impulse response (FIR) model is obtained when $na = nc = 0$. It can also be called a truncated weighting function model. Then

$$\begin{aligned} y(t) &= B(q^{-1})u(t) + e(t) \\ \theta &= (b_1 ... b_{nb}). \end{aligned} \tag{20}$$

- Another special case is when $nc = 0$. The model structure then becomes

$$\begin{aligned} A(q^{-1})y(t) &= B(q^{-1})u(t) + e(t) \\ \theta &= (a_1 ... a_{na} \quad b_1 ... b_{nb})^{T\cdot} \end{aligned} \tag{21}$$

This is sometimes called an ARX (controlled autoregressive) model. This structure can also be viewed as a linear regression,

$$y(t) = \varphi^T(t)\theta + e(t), \tag{22}$$

where

$$\varphi(t) = (-y(t-1) ... - y(t-na) \quad u(t-1) ... u(t-nb))^T . \tag{23}$$

There are some other ways than the ARMAX model to introduce polynomials for linear single-input single-output models. If $A(q^{-1})$ and $C(q^{-1})$ in Eq. (10) are constrained to coincide, the so-called *output error model*

$$y(t)=\frac{B(q^{-1})}{F(q^{-1})}u(t)+e(t) \tag{24}$$

is obtained, where all parameters are used to model the filter $G(q^{-1};\theta)$, while there is no description in the model of the disturbance.

The linear SISO models, such as ARX and ARMA models, can be extended to the multivariable case, but some complications occur in that there is no unique way to introduce free parameters. In case all matrix coefficients are left free, the model will lose uniqueness and it will then not be possible to identify the parameters, no matter what amount of data and what experimental condition is used.

Bibliography

Some reference for more extensive background material include

Anderson, B. D. O. and Moore, J. B. (1979): *Linear Optimal Filtering*,357pp. Prentice Hall, Englewood Cliffs, NJ. [Includes background and derivations for spectral factorization, innovations form, and optimal prediction].

Gevers, M. and Wertz, V. (1987): Techniques for selection of identifiable parametrizations for multivariable linear systems. In C T Leondes (ed.), *Control and Dynamic Systems*, vol 26; pp 35-86, System Identification and Adaptive Control – Advances in Theory and Application, Academic Press, New York. [Parameterization issues of linear multivariable systems].

Ljung, L. (1999): *System Identification*. *Theory for the User*, 2nd edition. 609pp.Prentice Hall. Upper Saddle River, NJ. [Description and analysis of prediction error and other identification methods].

Söderström, T. (2002): *Discrete-time Stochastic Systems*, 2nd ed., 375pp.Springer-Verlag, London, UK. [Includes background and derivations for spectral factorization, innovations form, and optimal prediction].

Söderström, T. and Stoica, P. (1989): *System Identification*,612pp.Prentice Hall International, Hemel Hempstead, UK[Description and analysis of prediction error and other identification methods].

Wahlberg. B. and. Ljung, L. (1986): Design variables for bias distribution in transfer function estimation. *IEEE Transactions on Automatic Control*, vol AC-31, pp 134-144. [Description of the effects of undermodeling]

Biographical Sketch

Torsten Söderström was born in Malmö, Sweden, in 1945. He received the MSc degree (`civilingenjör') in engineering physics in 1969 and the PhD degree in automatic control in 1973, both from Lund Institute of Technology, Lund, Sweden. He is a Fellow of IEEE.

In the period 1967-1974 he held various teaching positions at the Lund Institute of Technology. Since 1974, he has been at Department of Systems and Control, Uppsala University, Uppsala, Sweden, where he is a professor of automatic control.

Dr Söderström is the author or coauthor of more than 200 technical papers. His main research interests are in the fields of system identification, signal processing, and adaptive systems. He is the (co)author of four books: `Theory and Practice of Recursive Identification', MIT Press, 1983 (with L. Ljung), `The Instrumental Variable Methods for System Identification', Springer Verlag, 1983 (with P. Stoica), `System Identification', Prentice-Hall, 1989 (with P Stoica) and `Discrete-time Stochastic Systems', Prentice-Hall, 1994, second revised edition, Springer-Verlag, 2002. In 1981 he was, with coauthors, given an Automatica Prize Paper Award.

Within IFAC (International Federation of Automatic Control) he has served in several capacities including Automatica editor for the area of System Parameter Estimation since 1992, Council member 1996-2002, member of Executive Board and Chair of Awards Committee 1999-2002, IPC chairman of the SYSID'94 symposium, and guest associate editor or editor for three special issues of Automatica.