System Identification

ISY-Automatic Control
Linköping University

November 2012

Outline

Choosing a suitable model structure is prerequisite before estimation. The choice of model structure is based upon understanding of the physical systems. In system identification thee three most common models are

- Black-box models. This assumes that the system is unknown and model parameters adjustable without considering the physical background.
- Grey-box models. Assumes that part of the information about the dynamics or some physical parameters are known, and the model parameters might have some constraints.
- The user-defined models.

General linear models

A system can be described generally using the following equation, which is known as the general-linear polynomial model or the general-linear model

$$y(n) = q^{-k} G\left(q^{-1}, \theta\right) u(n) + H\left(q^{-1}, \theta\right) e(n) \qquad (1)$$

where $u(n)$ and $y(n)$ are the input and output, $e(n)$ is zero mean white noise, $G\left(q^{-1}, \theta\right)$ is the transfer function of the deterministic part of the system and $H\left(q^{-1}, \theta\right)$ is the transfer function of the stochastic part of the system.
Simpler models that are a subset of the general-linear model structure are possible, such as AR, ARX, ARMAX, Box-Jenkins, and output-error structures.

- **AR model.**Used in the generation of models where outputs are only dependent on previous outputs. Strictly speaking this model structure es the model for a signal, not a system.
- **ARX model.** Is the simplest model incorporating stimulus signal, the estimation of this model is the most efficient of the polynomial methods since the solution always satisfies the global minimum of the loss function. This model is preferable when the model order is high but disturbances are part of system dynamics.
- **ARMAX model.** This model structures includes disturbance dynamics, and are useful when dominating disturbances that enter early in the process are present. This model has more flexibility in the handling of disturbances modeling than the ARX model.

- **Box-Jenkins model.** This structure provides a complete model with disturbance properties modeled separately from system dynamics, it is useful when the disturbances enter late on the process.
- **Output-Error Model.** This model structure describes the system dynamics separately. No parameters are used for modeling the disturbance characteristics.

Some other models...

- **EWMA model.** The exponentially weighted moving average (EWMA) model is a particular case of the equation

$$\sigma_n^2 = \sum_{i=1}^{m} \alpha_i u_{n-i}^2$$

where the weights decrease exponentially as we move back through time. This approach has the attractive feature that relatively little data need to be stored. Is designed to track changes in the volatility.

- **ARIMA model.** The autoregressive integrated moving average (ARIMA) model is a generalization of an ARMA model. These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). They are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied to remove the non-stationarity.

- **ARCH model.** AutoRegressive Conditional Heteroskedasticity (ARCH) models are used to characterize and model observed time series. They are used whenever there is reason to believe that the terms will have a characteristic size, or variance. This models assume the variance of the current error term or innovation to be a function of the actual sizes of the previous time periods error terms: often the variance is related to the squares of the previous innovations. ARCH models are employed commonly in modeling financial time series that exhibit time-varying volatility clustering, i.e. periods of swings followed by periods of relative calm.

Prediction error

The essence of a model is its prediction aspect, and is also judged its performance in this respect. Let the prediction error given by a certain model $\mathscr{M}(\theta_\star)$ given by

$$\varepsilon(t,\theta_\star) = y(t) - \widehat{y}(t \mid \theta_\star) \tag{2}$$

When a data set Z^N is known, these errors can be computed for $t = 1,2,...,N$.
A good model is one that is good at predicting, that is, one that produces small prediction errors when applied to the observed data.
A guided principle for parameter estimation is:
Based on Z^N we can compute the prediction error $\varepsilon(t,\theta)$ using (2). At time $t = N$, select $\widehat{\theta}_N$ so that the prediction errors $\varepsilon\left(t,\widehat{\theta}_N\right)$, $t = 1,2,...,N$, become as small as possible.

Outline

Minimizing prediction errors

The prediction-error in (2) can be seen as a vector in $\mathbb{R}^N$. The size of this vector could be measured using any norm in $\mathbb{R}^N$. Let the prediction-error sequence be filtered through a stable linear filter

$$\varepsilon_F(t,\theta) = L(q)\varepsilon(t,\theta), 1 \leq t \leq N$$

Then use the following norm:

$$V_N\left(\theta, Z^N\right) = \frac{1}{N}\sum_{t=1}^{N} l\left(\varepsilon_F(t,\theta)\right) \tag{3}$$

where $l(.)$ is a scalar-valued function. The function $V_N\left(\theta, Z^N\right)$ is, for a given Z^N, a well-defined scalar-valued function of the model parameter. The estimate $\widehat{\theta}$ is then defined by minimization of (3):

$$\widehat{\theta}_N = \widehat{\theta}_N\left(Z^N\right) = \underset{\theta \varepsilon D_{\mathcal{M}}}{arg\,min} V_N\left(\theta, Z^N\right) \tag{4}$$

The term prediction-error identification methods (PEM) is used for the family of approaches that corresponds to (4). Particular methods with specific names are obtained as special cases of (4),depending on the choice of $l(\cdot)$, the choice of prefilter, the choice of model structure, and, in some cases, the choice of method which the minimization is realized.

Linear regressions

Linear regression model structures are very useful in describing basic linear and nonlinear systems. The linear regression employs a predictor

$$\hat{y}(t \mid \theta) = \varphi^T(\theta) + \mu(t) \tag{5}$$

that is linear in θ. Here φ is the vector of regressors, the regression vector. In (5), $\mu(t)$ is a known data-dependent vector. For notational simplicity $\mu(t) = 0$.

Least-squares criterion

With (5) the prediction error becomes

$$\varepsilon(t,\theta) = y(t) - \varphi^T(t)\,\theta \tag{6}$$

and the criterion function is

$$V_N\left(\theta, Z^N\right) = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{2}\left[y(t) - \varphi^T(t)\,\theta\right]^2 \tag{7}$$

This is the least-squares criterion for the linear regression (5), and it can be minimized analytically, which gives, provided the indicated inverse exists,

$$\widehat{\theta}_N^{LS} = arg\,min V_N\left(\theta, Z^N\right) = \left[\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\,\varphi^T(t)\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\,y(t) \tag{8}$$

the least-squares estimate (LSE).

Outline

Maximum Likelihood

In the are of statistical inference and system identification, we deal with the problem of extracting information from observations that could be unreliable, these are then described as realizations of stochastic variables. Suppose that the observations are represented by the random variable y^N that takes values in $\mathbb{R}^N$. The probability density function (PDE) is given by

$$P\left(y^N \varepsilon A\right) = \int_{x^N \varepsilon A} f_y\left(\theta; x^N\right) dx^N \tag{9}$$

Here θ is a d-dimensional parameter vector that describes properties of the observed variable and are unknown. The purpose of the observation is to estimate the vector θ using y^N, accomplished by $\widehat{\theta}\left(y^N\right)$. If the observed value of y^N is $y_\star^N$, then the resulting estimate is $\widehat{\theta}_\star = \widehat{\theta}\left(y_\star^N\right)$.

The probability that the observation indeed should take the value $y_\star^N$ is thus proportional to $f_y\left(\theta; y_\star^N\right)$. This function is called the likelihood function. A reasonable estimator of θ could then be to select it so that the observed event becomes "as likely as possible". That is

$$\widehat{\theta}_{ML}(y_\star^N) = arg\ \underset{\theta}{max} f_y\left(\theta; y_\star^N\right) \tag{10}$$

where the maximization is performed for fixed $y_\star^N$. This function is known as the maximum likelihood estimator (MLE).

Maximum a posteriori

For the Bayesian approach the parameter itself is thought of as a random variable, with this view we consider θ to be a random vector with a certain prior distribution. The observations y^N are correlated with this θ.
After the observations have been obtained, we then ask for the posterior PDF, from this, different estimates of θ can be determined. This is known as the maximum a posteriori estimate (MAP). The posterior PDF as a function of θ is thus proportional to the likelihood function multiplied by the prior PDF.
Often the prior PDF has an insignificant influence. The MAP estimate

$$\widehat{\theta}_{MAP}(y^N) = arg\max_{\theta} \left\{ f_y\left(\theta; y^N\right) \bullet g_\theta(\theta) \right\}$$

is close to the MLE.

Expectation Maximization

- The expectation maximization (EM) algorithm computes maximum likelihood (ML) estimates of unknown parameters θ in probabilistic models involving latent variables Z[1].
- The EM algorithm is an iterative method that alternates between computing a conditional expectation and solving a maximization problem, hence the name expectation maximization.
- The strategy of this algorithm is to separate the original ML problem into two linked problems, each of which is hopefully easier to solve than the original problem.

The key idea is to consider the joint log-likelihood function of both the observed variables Y and the latent variables Z

$$L_\theta(Z, Y) = log\, p_\theta(Z, Y) \tag{11}$$

and then assume that the latent variables Z were available to us. This algorithm has 2 steps. The first one is the expectation (E) step, that consists in computing the following

$$Q(\theta, \theta_k) \triangleq E\{log\, p_\theta(X, Y) \mid Y\} = \int log\, p_\theta(X, Y)\, p_{\theta_k}(X \mid Y)\, dX \tag{12}$$

and then the maximization (M) step that amounts to solving the following problem

$$\theta_{k+1} = arg\, max_\theta \supset (\theta, \theta_k) \tag{13}$$

Outline

Instrumental Variables

Consider again the linear regression model $\widehat{y}(t \mid \theta) = \varphi^T(t)\theta$. The LS estimate can also be expressed as

$$\widehat{\theta}_N^{LS} = sol\left\{\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\left[y(t) - \varphi^T(t)\theta\right] = 0\right\} \tag{14}$$

Now suppose that the data actually can be described by

$$y(t) = \varphi^T(t)\theta_0 + v_0(t) \tag{15}$$

The LSE $\widehat{\theta}_N$ will not tend to θ_0 in typical cases, because of the correlation between $v_0(t)$ and $\varphi(t)$. Let us try instead a general correlation vector $\zeta(t)$ in (11).This is called an instrumental-variable method (IV).

The elements of are then called instruments or instrumental variables. This gives

$$\widehat{\theta}_N^{IV} = sol\left\{\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\left[y(t)-\varphi^T(t)\theta\right]=0\right\} \tag{16}$$

or

$$\widehat{\theta}_N^{IV} = \left[\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\varphi^T(t)\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\zeta(t)y(t) \tag{17}$$

provided the indicated inverse exists.
For this method we could say that the instruments must be correlated with the regression variables but uncorrelated with the noise.

An example

Consider the problem of estimating the variance of a variable X from m observations on X when the underlying distribution is normal with zero mean. The observations are $u_1, u_2, ..., u_m$. The variance is v. The likelihood of u_i being observed is defined as the probability density function when $X = u_i$

$$\frac{1}{\sqrt{2\pi v}} exp\left(\frac{-u_i^2}{2v}\right)$$

the likelihood of m observations occurring in the order in which they are observed is

$$\prod_{i=1}^{m}\left[\frac{1}{\sqrt{2\pi v}} exp\left(\frac{-u_i^2}{2v}\right)\right]$$

Using the MLE, the best estimate of v is the value that maximizes this expression.

Taking logarithms of the previous expression and ignoring constant multiplicative factors, we wish to maximize

$$\sum_{i=1}^{m}\left[-ln(\nu)-\frac{u_i^2}{\nu}\right]$$

or

$$-mln(\nu)-\sum_{i=1}^{m}\frac{u_i^2}{\nu}$$

differentiating this with respect to ν and setting the resulting equation to zero, the MLE estimator of ν is

$$\frac{1}{m}\sum_{i=1}^{m}u_i^m$$

For Further Reading I

Ljung L.
System Identification. Theory for the user.
Prentice Hall, 1987.

Schön T.
An Explanation of the Expectation Maximization Algorithm
Technical report from Automatic Control at LiU, 2009.