System Identification

9

System identification is concerned with the estimation of a system on the basis of observed data. This involves specification of the model structure, estimation of the unknown model parameters, and validation of the resulting model. Least squares and maximum likelihood methods are discussed, for stationary processes (without inputs) and for input-output systems.

9.1 Identification

In the foregoing chapters we always assumed that the system is known to us, and we considered the representation, regulation, and prediction of linear systems with given parameters. In most practical applications the system is not known and has to be estimated from the available information. This is called the identification problem. The identification method will depend on the intended model use, as this determines what aspects of the system are of relevance. The three main choices in system identification are the following.

(i) *Data* In some situations it is possible to generate a large amount of reliable data by carefully designed experiments. In other situations the possibilities to obtain data are much more limited and it is not possible to control for external factors that influence the outcomes. That is, the magnitude of outside disturbances ('noise') may differ widely from one application to another.

(ii) *Model Class* A model describes relations between the observed variables. For practical purposes the less important aspects are neglected to obtain sufficiently simple models. The identified model should be validated to test whether the imposed simplifications are acceptable.

C. Heij et al., *Introduction to Mathematical Systems Theory*,
https://doi.org/10.1007/978-3-030-59654-5_9

(iii) *Criterion* The criterion reflects the objectives of the modeller. It expresses the usefulness of models in representing the observed data.

In practice, system identification often involves several runs of the empirical cycle which consists of the specification of the problem, the estimation of a model by optimization of the criterion, the validation of the resulting model, and possible adjustments that may follow from this validation.

In the following we restrict our attention to linear systems, quadratic criteria and data that consists of observed time series of the system variables. The advantage of this linear quadratic framework is that it leads to relatively simple identification algorithms. Further, the ideas and concepts for these methods form the basis for more advanced approaches.

Models are simplifications of reality and therefore they involve errors. It is often assumed that the data can be decomposed into two parts, a systematic part (related to the underlying system) and a disturbance part that reflects unmodelled aspects of the system. By assuming that the disturbances are random variables, the statistical properties of identification methods can be evaluated. In particular, one considers the properties of unbiasedness, efficiency, and consistency. Let θ denote the unknown system parameters, and let $\hat{\theta}$ be an estimator of θ based on the observed data. Because the data are influenced by the random disturbances, the estimator $\hat{\theta}$ is also a random variable. It is called an *unbiased estimator* if $E(\hat{\theta}) = \theta$, and it is called an *efficient estimator* in a class of estimators if it minimizes the variance $\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))(\hat{\theta} - E(\hat{\theta}))^T$, that is, if for every other estimator $\tilde{\theta}$ in this class $\text{var}(\tilde{\theta}) - \text{var}(\hat{\theta})$ is a positive semidefinite matrix. To define consistency, let $\hat{\theta}_N$ denote the estimator based on data that are observed on a time interval of length N. The estimator is called (*weakly*) *consistent* if, for every $\delta > 0$, there holds

$$\lim_{N\to\infty} P(\|\hat{\theta}_N - \theta\| \geq \delta) = 0 \tag{9.1}$$

where $\|\cdot\|$ denotes the Euclidean norm. This is also written as $\text{plim}(\hat{\theta}_N) = \theta$. Hereby it is assumed that the system under investigation belongs to the model class, but this can be generalized to the situation where θ is the optimal (but not perfectly correct) model within the model class.

9.2 Regression Models

In this section we consider single input, single output systems with a finite impulse response (FIR), that is,

$$y(t) = \beta_1 u(t-1) + \cdots + \beta_k u(t-k) + \varepsilon(t) \tag{9.2}$$

We assume that y is observed for $t = 1, \ldots, N$, and u for $t = 1 - k, \ldots, N - 1$ with $N \geq k$. Let $x(t) := \big(u(t-1), \ldots, u(t-k)\big)^T$ and let $y = \big(y(1), \cdots, y(N)\big)^T$, $X = \big(x(1), \ldots, x)^T$, $\varepsilon = \big(\varepsilon(1), \ldots, \varepsilon(N)\big)^T$ and $\beta = (\beta_1, \ldots, \beta_k)^T$. Then (9.2) can be written as the regression model

$$y = X\beta + \varepsilon. \tag{9.3}$$

In the sequel, whenever necessary, we shall write X_N instead of X to emphasize the dependence of X on N.

From the data, y and X, we have to estimate the parameters β. The least squares estimator $\hat{\beta}$ minimizes the sum of squared errors

$$\sum_{t=1}^{N} \varepsilon^2(t) = \|\varepsilon\|^2 = \|y - X\beta\|^2.$$

This is obtained by projecting y onto the column space of X, so that
$X^T(y - X\hat{\beta}) = 0$. Assuming that $\text{rank}(X) = k$, the solution is given by

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{9.4}$$

In order to investigate under which conditions this is a good estimator, we make the following assumptions.

Assumptions
The data satisfy the relation $y = X\beta + \varepsilon$, where

A1 all entries of the matrix X are non-random, and $\text{rank}(X) = k$;
A2 all entries of the (unobserved) disturbance vector ε are outcomes of random variables with $E(\varepsilon) = 0$, $E(\varepsilon^2(t)) = \sigma^2$ (equal variance), and $E(\varepsilon(t)\varepsilon(s)) = 0$ for all $t \neq s$ (no serial correlation).

Definition 9.2.1 We call an estimator *linear* if it is of the form $\tilde{\beta} = Ay$, with A a non-random matrix, and it is called a *best linear unbiased estimator* (BLUE) if it is unbiased with minimal variance in the class of all linear unbiased estimators.

The following result is called the *Gauss-Markov theorem*.

Theorem 9.2.2 *Under assumptions A1 and A2, the least squares estimator* (9.4) *is BLUE with* $\text{var}(\hat{\beta}) = \sigma^2(X^T X)^{-1}$. *A sufficient condition for consistency is that*

$$\lim_{N\to\infty} \lambda_{min}(X_N^T X_N) = \infty,$$

where X_N is the regressor matrix in (9.3) for the first N observations and λ_{min} denotes the smallest eigenvalue.

Proof It follows from (9.3) and (9.4) that $\hat{\beta} = \beta + (X^T X)^{-1} X^T \varepsilon$. As X is non-random, $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \sigma^2 I$, it follows that $E(\hat{\beta}) = \beta$ and

$$\text{var}(\hat{\beta}) = (X^T X)^{-1} X^T \text{var}(\varepsilon) X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Let $\tilde{\beta} = Ay$ be another unbiased estimator and define $\Delta = A - (X^T X)^{-1} X^T$. Unbiasedness requires that $E(\tilde{\beta}) = AX\beta = \beta$ for every β, so that $AX = I$ and $\Delta X = 0$. Then $\tilde{\beta} - E\tilde{\beta} = A(X\beta + \varepsilon) - \beta = A\varepsilon$ and

$$\begin{aligned}\text{var}(\tilde{\beta}) &= E(\tilde{\beta} - E\tilde{\beta})(\tilde{\beta} - E\tilde{\beta})^T = E(A\varepsilon\varepsilon^T A^T) = \sigma^2 A A^T \\ &= \sigma^2(\Delta\Delta^T + (X^T X)^{-1}) = \sigma^2 \Delta\Delta^T + \text{var}(\hat{\beta}).\end{aligned}$$

As $\Delta\Delta^T$ is positive semidefinite this shows that $\hat{\beta}$ is BLUE.

From now on we emphasize that $X = X_N$ and denote $\hat{\beta}$ by $\hat{\beta}_N$.

To prove consistency we use the Markov inequality, that is, for every random variable z and every $c > 0$ there holds $E(z^2) \geq c^2 P(\mid z \mid \geq c)$ so that $P(\mid z \mid \geq c) \leq c^{-2} E(z^2)$. It then follows that for every $\delta > 0$

$$\begin{aligned}&P(\|\hat{\beta}_N - \beta\| \geq \delta) \leq P(\mid \hat{\beta}_{n,i} - \beta_i \mid \geq k^{-\frac{1}{2}}\delta \text{ for some } i = 1, \cdots, k) \leq \\ &\leq k\delta^{-2} E(\hat{\beta}_{N,i} - \beta_i)^2 = k\delta^{-2}\text{var}(\hat{\beta}_{N,i}) \leq k\delta^{-2}\sigma^2 \lambda_{max}\{(X_N^T X_N)^{-1}\} = \\ &= k\delta^{-2}\sigma^2\{\lambda_{min}(X_N^T X_N)\}^{-1}\end{aligned}$$

and this converges to zero for $N \to \infty$, by assumption. □

Returning to the FIR system (9.2), assumptions A1 and A2 mean that the input is not random but the output is random. This may be relevant in experimental situations where the input is controlled. However, often the input will be affected by uncertain factors that fall outside the scope of the model. The above results remain asymptotically valid for random inputs, provided some conditions are satisfied. We restrict the attention to consistency, and replace assumption A1 by the following.

A1* The matrix X is random and such that $\text{plim}(\frac{1}{N} X_N^T X_N) = Q$ exists with Q invertible (sufficiency of excitation).

For the FIR system (9.2) there holds $\frac{1}{N} X_N^T X_N = \frac{1}{N}\sum_{t=1}^N x(t)^T x(t)$, where $x(t) = \big(u(t-1), \ldots, u(t-k)\big)$, so that Q corresponds to the covariance matrix of the input and its lags. The excitation condition basically means that the input satisfies no polynomial equations and that it does not die out when $N \to \infty$.

Theorem 9.2.3 *Under assumptions A1* and A2, the least squares estimator is consistent if and only if* $\text{plim}(\frac{1}{N}\sum_{t=1}^{N} x(t)^T \varepsilon(t)) = 0$ *(orthogonality condition).*

Proof The least squares estimator is $\hat{\beta}_N = \beta + (\frac{1}{N}X_N^T X_N)^{-1}(\frac{1}{N}X_N^T \varepsilon)$ where $\frac{1}{N}X_N^T\varepsilon = \frac{1}{N}\sum_{t=1}^{N} x(t)^T\varepsilon(t)$. The definition of convergence in probability gives that if $\text{plim}(a_n) = a$ and f is a continuous function, then $\text{plim}(f(a_n)) = f(a)$. Therefore $\text{plim}(\hat{\beta}_N) = \beta + Q^{-1}\text{plim}(\frac{1}{N}X_N^T\varepsilon)$, which proves the result. □

The orthogonality condition essentially requires that the regressor variables $x(t)$ show no contemporaneous correlation with the error term $\varepsilon(t)$. For the FIR system this means that the output error in (9.2) is uncorrelated with the past inputs.

Many time series that are observed in practice show trends and seasonal variation. The modelling of trends and seasonals is discussed in the next chapter. In the current chapter we will either assume that the data are stationary, which can sometimes be achieved by appropriate data transformations, or that the model explicitly includes variables for the nonstationary part.

9.3 Maximum Likelihood

Stochastic models assign (relative) probabilities to the observations of the system variables. Suppose that the model class consists of a set of probability densities $\{p_\theta, \theta \in \Theta\}$, where $\theta \in \Theta$ is the vector of unknown parameters. If the data consists of q time series that are observed on a time interval of length N, then p_θ is a probability density on $(\mathbb{R}^q)^N$. The maximum likelihood method chooses the model that assigns the highest probability to the observed data. If we denote the data by $w \in (\mathbb{R}^q)^N$, then this means that the likelihood function $L(\theta) := p_\theta(w)$ is maximized over the parameter set Θ.

Maximum likelihood estimation (ML) requires that the probability distribution is specified as an explicit function of the parameters θ. As an example, we consider the regression model (9.3) $y = X\beta + \varepsilon$. In this case, the parameters θ are given by $(\beta^T, \sigma^2)^T$. We extend assumption A2 as follows.

A2* The disturbance vector ε has the multivariate normal distribution with mean $E(\varepsilon) = 0$ and covariance matrix $E(\varepsilon\varepsilon^T) = \sigma^2 I$.

Theorem 9.3.1 *Under assumptions A1 and A2*, the maximum likelihood estimators in the regression model* (9.3) *are given by* $\hat{\beta} = (X^TX)^{-1}X^Ty$ *and* $\hat{\sigma}^2 = \frac{1}{N}(y - X\hat{\beta})^T(y - X\hat{\beta})$.

Proof Let $\theta = (\beta^T, \sigma^2)^T$ denote the vector of the model parameters. As $\varepsilon = y - X\beta$ has the normal distribution, the likelihood function is given by

$$L(\beta, \sigma^2) = p_\theta(y, X) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\{-(2\sigma^2)^{-1}(y - X\beta)^T(y - X\beta)\} \qquad (9.5)$$

As the logarithm is a monotonic function, maximization of $L(\beta, \sigma^2)$ is equivalent to maximization of

$$\frac{2}{N}\log L(\beta, \sigma^2) = -\log(2\pi) - \log(\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{N}(y - X\beta)^T(y - X\beta).$$

It follows that the maximum is obtained for $\hat{\beta} = (X^T X)^{-1}X^T y$ and that $\hat{\sigma}^2 = \frac{1}{N}(y - X\hat{\beta})^T(y - X\hat{\beta})$. □

Theorem 9.3.2 *Under assumptions A1 and A2*, the least squares estimator $\hat{\beta}$ in* (9.4) *is minimum variance unbiased, that is, it is unbiased and if $\tilde{\beta}$ is another unbiased estimator then* $\text{var}(\tilde{\beta}) - \text{var}(\hat{\beta})$ *is positive semidefinite.*

Proof Again, let $\theta = (\beta^T, \sigma^2)^T$ denote the model parameters. The Cramer-Rao theorem states that every unbiased estimator $\hat{\theta}$ has a covariance matrix $\text{var}(\hat{\theta}) \geq [-E(\frac{\partial^2 \log L}{\partial\theta\partial\theta^T})]^{-1}$, see [40]. It follows by direct calculation from (9.5) that in this case the lower bound is a block-diagonal matrix with blocks $\sigma^{-2}(X^T X)$ and $(2\sigma^4)^{-1}N$. This implies that for every unbiased estimator there holds $\text{var}(\tilde{\beta}) \geq \sigma^2(X^T X)^{-1} = \text{var}(\hat{\beta})$, see Theorem 9.2.2. □

Under very general conditions, maximum likelihood estimators have optimal asymptotic properties, provided that the model is correctly specified. That is, if the data are generated by a probability distribution p_{θ^0}, with $\theta^0 \in \Theta$, and $\hat{\theta}_N$ is the ML estimate based on N observations, then under very general conditions there holds that

(i) $\hat{\theta}_N$ is consistent, that is, $\text{plim}(\hat{\theta}_N) = \theta^0$;
(ii) $\hat{\theta}_N$ is asymptotically efficient in the class of all consistent estimators, that is, $\lim_{N\to\infty} N(\text{var}(\tilde{\theta}_N) - \text{var}(\hat{\theta}_N))$ is positive semidefinite for every consistent estimator $\tilde{\theta}$;
(iii) $\hat{\theta}_N$ has an asymptotic normal distribution, in the sense that $\sqrt{N}(\hat{\theta}_N - \theta^0)$ converges to a normal distribution with mean zero and covariance matrix $[-E(\frac{\partial^2 \log L}{\partial\theta\partial\theta^T})]^{-1}$.

We refer to , e.g., [25] for a proof of this result. From a computational point of view, ML estimation requires the maximization of the likelihood function or equivalently, of its logarithm, both of which are functions of several real variables. The first order conditions will in general consist of a set of nonlinear equations in θ that can be solved by numerical methods. Such methods differ in the choice of initial estimates, search strategies, and convergence criteria. The Newton-Raphson method consists of an iterative linearization of the stationarity condition for a maximum. Consider this for the maximization of the logarithm of the likelihood functions. If $\hat{\theta}_i$ is the current estimate, $G_i = \frac{\partial log L(\theta)}{\partial\theta}$ the gradient and $H_i = \frac{\partial^2 \log L}{\partial\theta\partial\theta^T}$ the Hessian in $\hat{\theta}_i$, then locally around $\hat{\theta}_i$ there holds

$\frac{\partial \log L(\theta)}{\partial \theta} \approx G_i + H_i(\theta - \hat{\theta}_i)$ by Taylor's formula. This motivates the iterations

$$\hat{\theta}_{i+1} = \hat{\theta}_i - H_i^{-1} G_i \tag{9.6}$$

A possible disadvantage is that this requires the computation and inversion of the Hessian matrix. For nonlinear regression models of the form

$$y(t) = f(x(t), \theta) + \varepsilon(t) \tag{9.7}$$

one could use the Gauss-Newton method for the minimization of $\sum_{t=1}^{N} \varepsilon^2(t)$ as an alternative. This corresponds to maximum likelihood if the disturbances satisfy assumption A2*. Here $x(t)$ is the vector of regressors at time t, and f is a nonlinear function of the model parameters θ. If $\hat{\theta}_i$ is the current estimate, then the model (9.7) is linearized by $f(x, \theta) \approx f(x, \hat{\theta}_i) + x_i^T(\theta - \hat{\theta}_i)$, where $x_i = \frac{\partial}{\partial \theta} f(x, \theta)$ is the gradient evaluated at $(x, \hat{\theta}_i)$. The linearized model gives $\varepsilon(t) = y(t) - f(x(t), \theta) \approx y(t) - f(x(t), \hat{\theta}_i) - x_i^T(t)(\theta - \hat{\theta}_i) = \varepsilon_i(t) - x_i^T(t)(\theta - \hat{\theta}_i)$, where $\varepsilon_i(t)$ denotes the residuals of (9.7) for the estimate $\hat{\theta}_i$ and $x_i(t)$ is the gradient of f at $(x(t), \hat{\theta}_i)$. The corresponding approximation of the criterion function gives $\sum_{t=1}^{N} \varepsilon^2(t) \approx \sum_{t=1}^{N} \{\varepsilon_i(t) - x_i^T(t)(\theta - \hat{\theta}_i)\}^2$. This is a least squares problem with estimate $\hat{\theta}_{i+1} = (X_i^T X_i)^{-1} X_i^T (\varepsilon_i + X_i \hat{\theta}_i)$, that is

$$\hat{\theta}_{i+1} = \hat{\theta}_i + (X_i^T X_i)^{-1} X_i^T \varepsilon_i \tag{9.8}$$

Here X_i is the matrix with N rows consisting of the gradients $x_i(t)$, $t = 1, \cdots, N$, and ε_i is the $N \times 1$ vector with the residuals for $\hat{\theta}_i$.

9.4 Estimation of Autoregressive Models

In this section, we suppose that the data consists of observations of a single output variable $y(t)$, observed for $t = 1, \cdots, N$, and generated by an autoregressive model

$$y(t) = \alpha_1 y(t-1) + \cdots + \alpha_p y(t-p) + \varepsilon(t). \tag{9.9}$$

Here ε is a white noise process with mean zero, variance σ^2, and finite fourth order moments, so that assumption A2 is satisfied. We assume that this model is causal, that is, that the polynomial $1 - \sum_{i=1}^{p} \alpha_i z^{-i}$ has all its roots inside the unit disc. Moreover, we assume that p is known and correctly specified. In Sect. 9.6.1 we shall discuss methods to estimate the lag order p from the data.

Theorem 9.4.1 *The least squares estimator of* $(\alpha_1, \cdots, \alpha_p)$ *in a causal autoregressive model* (9.9) *is consistent.*

Outline of Proof According to Theorem 9.2.3, it suffices to prove that assumption A1* is satisfied and that $\text{plim}(\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t)y(t-i)) = 0$ for $i = 1, \cdots, p$. As was discussed in Sect. 6.3, stationarity implies that $y(t)$ can be written as a function of the past disturbances $\{\varepsilon(s), s \leq t\}$. Therefore $E(\varepsilon(t)y(t-i)) = 0$ for all t and $i = 1, \cdots, p$, so that $\varepsilon(t)$ is uncorrelated with all the regressors in (9.9). This means that $\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t)y(t-i)$ is the sample mean of N mutually uncorrelated terms with mean 0 and constant variance $E(\varepsilon(t)y(t-i))^2 < \infty$, because ε has finite fourth order moments. The weak law of large numbers implies that

$$\text{plim}(\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t)y(t-i)) = 0.$$

As concerns assumption A1*, $\frac{1}{N}X_N^T X_N$ is a $p \times p$ matrix with (i, j)-th element $\frac{1}{N}\sum_{t=1}^{N} y(t-i)y(t-j)$. Under the above conditions the process y can be shown to be ergodic. The proof requires a generalized law of large numbers for the sample mean of N correlated terms (but with exponentially decaying correlation between $y(t-i)y(t-j)$ and $y(t-i+k)y(t-j+k)$ for $k \to \infty$). Ergodicity implies that the matrix Q in assumption A1* exists, and that $Q_{ij} = E(y(t-i)y(t-j))$. Further Q is invertible, because otherwise there would exist $a \in \mathbb{R}^p$ such that $a^T Qa = \text{var}(\sum_{i=1}^{p} a_i y(t-i)) = 0$ which contradicts that the autoregressive process (9.9) has no perfectly predictable component. □

In the model (9.9) the observations have mean $Ey(t) = 0$. In practice, one may add regressors to take care of, for example, non-zero mean and trends, so that

$$y(t) = \mu_1 + \mu_2 T(t) + \alpha_1 y(t-1) + \cdots + \alpha_p y(t-p) + \varepsilon(t). \tag{9.10}$$

Least squares is also consistent for this model under the conditions of Theorem 9.4.1.

Theorem 9.4.2 *If in the autoregressive model* (9.9) *the noise* ε *satisfies assumption* A2* *(normality), then the least squares estimator is consistent, asymptotically efficient, and asymptotically normally distributed.*

Proof It is sufficient to prove that under these conditions least squares is asymptotically equivalent to maximum likelihood. The likelihood function of (9.9) can be written, by conditioning, as

$$\begin{aligned} L(\alpha_1, \cdots, \alpha_p) &= p(y(1), \cdots, y(N)) \\ &= p(y(1), \cdots, y(p))\Pi_{t=p+1}^{N} p(y(t) \mid y(1), \cdots, y(t-1)) \\ &= p(y(1), \cdots, y(p))\Pi_{t=p+1}^{N} p(y(t) \mid y(t-p), \cdots, y(t-1)) \\ &= p(y(1), \cdots, y(p))\Pi_{t=p+1}^{N} p(\varepsilon(t)). \end{aligned}$$

As $p(\varepsilon(t)) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-(2\sigma^2)^{-1}\varepsilon(t)^2\}$ this gives

$$\frac{1}{N}\log L = \frac{1}{N}\log(p(y(1),\cdots,y(p))) + \frac{1}{N}\sum_{t=p+1}^{N}\log p(\varepsilon(t))$$

$$= \frac{1}{N}\log(p(y(1),\cdots,y(p))) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(2\sigma^2)^{-1}}{N}\sum_{t=p+1}^{N}\varepsilon(t)^2.$$

Apart from the first term, that vanishes for $N \to \infty$, this shows that the ML estimates of $\alpha_1, \cdots, \alpha_p$ are obtained by minimizing $\sum_{t=p+1}^{N}\varepsilon(t)^2$. □

There is a close connection between least squares and the so-called *Yule-Walker equations*. As $E(\varepsilon(t)y(t-i)) = 0$ for $i = 1, \cdots, p$, it follows from (9.9) that the autocovariances $R(k) = E(y(t)y(t-k))$ of the process y satisfy

$$\begin{pmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{pmatrix} = \begin{pmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}. \tag{9.11}$$

If we replace $R(k)$ by $\hat{R}(k) = \frac{1}{N}\sum_{t=k+1}^{N} y(t)y(t-k)$ then (9.11) can be solved for the parameters $\alpha_i, i = 1, \cdots, p$. For numerical reasons, the autocovariances are often scaled by using the correlations $\hat{\rho}(k) = \hat{R}(k)/\hat{R}(0)$ in (9.11) instead of $\hat{R}(k)$. That is, one considers estimates $\hat{\alpha}_j$ obtained by solving the following set of linear equations:

$$\begin{pmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \\ \vdots \\ \vdots \\ \hat{\rho}(p) \end{pmatrix} = \begin{pmatrix} 1 & \hat{\rho}(1) & \cdots & \cdots & \hat{\rho}(p-1) \\ \hat{\rho}(1) & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \hat{\rho}(1) \\ \hat{\rho}(p-1) & \cdots & \cdots & \hat{\rho}(1) & 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \vdots \\ \hat{\alpha}_p \end{pmatrix}. \tag{9.12}$$

The structure of the matrix in the right hand side of this equation is a very special one: it is symmetric positive definite, but also it is a Toeplitz matrix: along diagonals the same entry occurs. Fast methods to solve sets of equations of this kind for $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$ are important, in particular in cases where p is large. One such fast algorithm is known as the Levinson algorithm; it requires considerably fewer numerical operations than the $O(p^3)$ operations needed for Gaussian elimination. See, e.g., [20].

To discuss the estimation of σ^2 resulting from the estimates for the α_j we use the fact that

$$\varepsilon(t) \approx \hat{\varepsilon}(t) = y(t) - \hat{\alpha}_1 y(t-1) - \cdots - \hat{\alpha}_p y(t-p).$$

Note that $\sigma^2 = E(\varepsilon(t)^2) = E(\varepsilon(t)y(t))$. Replacing in the latter formula $\varepsilon(t)$ by $\hat{\varepsilon}(t)$ we arrive at the following estimate $\hat{\sigma}^2$ for σ^2:

$$\hat{\sigma}^2 = E(\hat{\varepsilon}(t)y(t)) = \hat{R}(0) - \hat{\alpha}_1\hat{R}(1) - \cdots - \hat{\alpha}_p\hat{R}(p).$$

One can check that the estimates resulting from solving (9.12) are approximately equal to the least squares estimates (where the summations run from $t = p + 1$ to N instead of from $t = k + 1$ to N).

Next we consider autoregressive models with inputs, that is,

$$y(t) = \sum_{i=1}^{p} \alpha_i y(t-i) + \sum_{i=0}^{q} \beta_i u(t-i) + \varepsilon(t) \tag{9.13}$$

Such a model is also called an ARX model, that is, an autoregressive model with exogenous variables. We assume that $\sum_{i=1}^{p} \alpha_i y(t-i) + \sum_{i=0}^{q} \beta_i u(t-i)$ is the optimal linear predictor of $y(t)$, in the sense that it minimizes the mean squared prediction error $E(y(t) - \hat{y}(t))^2$ over the class of all linear predictors of the type $\hat{y}(t) = \sum_{i\geq 0}(a_i y(t-i-1)+b_i u(t-i))$. Optimality implies that $E((y(t)-\hat{y}(t))\hat{y}(t)) = 0$, so that $E(\varepsilon(t)y(t-i)) = 0$ for all $i \geq 1$ and $E(\varepsilon(t)u(t-i)) = 0$ for all $i \geq 0$. Further we assume that the uncontrolled system with input $u(t) = 0$ is causal, that is, that $1 - \sum_{i=1}^{p} \alpha_i z^{-i}$ has all its roots inside the unit disc. We use the notation $\theta = (\alpha_1, \cdots, \alpha_p, \beta_0, \cdots, \beta_q)^T$, $x(t) = (y(t-1), \cdots, y(t-p), u(t), u(t-1), \cdots, u(t-q))^T$, and

$$Q_N = \begin{pmatrix} Q_N(yy) & Q_N(yu) \\ Q_N(uy) & Q_N(uu) \end{pmatrix} = \frac{1}{N}\sum_{t=m}^{N} x(t)x(t)^T$$

where $m = \max\{p, q\}$. So $[Q_N(yy)]_{ij} = \frac{1}{N}\sum_{t=m}^{N} y(t-i)y(t-j) = \hat{R}_y(i-j), i, j = 1, \cdots, p$, are the sample autocovariances of the output, and similarly for the other entries of the matrix Q_N.

Theorem 9.4.3 *Under the above conditions, the least squares estimators of the parameters in the ARX system* (9.13) *are consistent if the inputs are sufficiently excited in the sense that* $\text{plim} Q_N(uu) = Q(uu)$ *exists and is invertible.*

Details of the proof fall outside the scope of this book, we refer to [21]. The idea is similar to the proof of Theorem 9.4.1. That is, the least squares estimator is given by

$\hat{\theta}_N = \theta + Q_N^{-1}\delta_N$ where $\delta_N = \frac{1}{N}\sum_{t=m+1}^{N}\varepsilon(t)x(t)$. As $\text{plim}Q_N(uu)$ exists and the system (9.13) is causal, it follows that also $\text{plim}Q_N(yy) = Q(yy)$ and $\text{plim}Q_N(yu) = Q(yu)$ exist. Further, $Q = \text{plim}Q_N$ is invertible, because otherwise there would exist $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^{q+1}$ such that $(a^T, b^T)Q(a^T, b^T)^T = \text{var}(\sum_{i=1}^{p} a_i y(t-i) + \sum_{i=0}^{q} b_i u(t-i)) = 0$. Because $Q(uu)$ is invertible, $a_i \neq 0$ for at least one $i = 1, \cdots, p$, and this contradicts the fact that $y(t)$ is not perfectly predictable from the observations $\{y(s-1), u(s), s \leq t\}$. Therefore, $\text{plim}(\hat{\theta}_N) = \theta + Q^{-1}\text{plim}(\delta_N)$, and $\text{plim}(\delta_N) = 0$. This orthogonality condition again follows from a weak law of large numbers.

Note that this result does not require that the input is deterministic. It may, for instance, be generated by feedback, where $u(t)$ depends on the past outputs $\{y(s), s \leq t-1\}$. However, the input $u(t)$ may not depend on the current output $y(t)$, as in this case the orthogonality condition $E(\varepsilon(t)u(t)) = 0$ would be violated. The input condition stated in Theorem 9.4.3 can be weakened, but some persistency of excitation is needed.

In the foregoing we restricted our attention to systems (9.9) with one output and (9.13) with one input and one output. Similar results hold true for multivariate systems, with multiple inputs and outputs.

9.5 Estimation of ARMAX Models

In the foregoing section it was assumed that the disturbances $\varepsilon(t)$ in (9.9) and (9.13) are white noise. If the disturbances are correlated over time then this indicates that the dynamic specification of the model is not correct. This can be repaired by increasing the lag orders of the model, but this may lead to a large number of parameters. It may then be preferable to estimate more parsimonious models. For example, for single-input, single-output systems one can use ARMAX models defined by

$$y(t) = \sum_{i=1}^{p}\alpha_i y(t-i) + \sum_{i=0}^{q}\beta_i u(t-i) + \varepsilon(t) + \sum_{i=1}^{r}\gamma_i\varepsilon(t-i) \tag{9.14}$$

If the inputs are $u(t) = 0$, then this is an ARMA model. We assume that this model is coprime, causal and invertible, i.e., the equations $1 - \sum_{i=1}^{p}\alpha_i z^{-i} = 0$ and $1 + \sum_{i=1}^{r}\gamma_i z^{-i} = 0$ have all their solutions in $| z |< 1$ and the equations have no common solutions. The white noise process $\varepsilon(t)$ then has the interpretation of the one-step ahead prediction errors, see Sect. 6.3.

Theorem 9.5.1 *For an ARMAX system* (9.14) *with* $p \neq 0$ *and* $r \neq 0$, *the least squares estimate in the regression model* (9.13) *is in general not consistent.*

Proof The disturbances in the model (9.13) are given by $\varepsilon(t) + \sum_{i=1}^{r}\gamma_i\varepsilon(t-i)$. If $p \neq 0 \neq r$, then these are in general correlated with the output regressors in (9.13).

Therefore the orthogonality condition is violated, and it follows from Theorem 9.2.3 that least squares is not consistent.

As a simple example, consider the ARMA(1,1) model $y(t) = \alpha y(t-1) + \varepsilon(t) + \gamma\varepsilon(t-1)$ with $\alpha \neq 0 \neq \gamma$ and $|\alpha| < 1, |\gamma| < 1$. The least squares estimate of α is given by $\hat{\alpha}_N = (\sum_{t=2}^{N} y(t)y(t-1))/(\sum_{t=2}^{N} y^2(t-1))$. From this it follows that $\text{plim}(\hat{\alpha}_N) = \alpha + \gamma\sigma^2/\text{var}(y(t))$. This is inconsistent if $\gamma \neq 0$. □

Consistent estimators may be obtained by using so-called *instrumental variables*. We formulate this in terms of the regression model (9.3), with $\text{plim}(\frac{1}{N}X_N^T\varepsilon_N) \neq 0$ where X_N is the $N \times k$ regressor matrix and ε_N the $N \times 1$ disturbance vector for sample size N. The variables $z_i(t), i = 1, \cdots, l$, are called *instruments* if the following conditions are satisfied, where Z_N denotes the $N \times l$ matrix with elements $z_i(t)$.

$$\begin{gathered}\text{plim}(\frac{1}{N}Z_N^T\varepsilon_N) = 0,\ \text{plim}(\frac{1}{N}Z_N^TZ_N) = Q_{zz},\ \text{plim}(\frac{1}{N}Z_N^TX_N) = Q_{zx} \\ \text{rank}(Q_{zz}) = l,\ \text{rank}(Q_{zx}) = k.\end{gathered} \tag{9.15}$$

The idea is to replace the regressors X_N by the instruments Z_N, because they satisfy the orthogonality condition. In order to approximate X_N as well as possible, they are regressed on Z_N. Therefore, the instrumental variables estimator $\hat{\theta}_{IV}$ is defined by the following two steps. First regress X_N on Z_N, with fitted values $\hat{X}_N = Z_N(Z_N^TZ_N)^{-1}Z_N^TX_N$, and then regress y on $\hat{X}_N$. Let $P_N = Z_N(Z_N^TZ_N)^{-1}Z_N^T$ be the projection operator on the column space of Z_N,, then

$$\hat{\theta}_{IV} = (\hat{X}_N^T\hat{X}_N)^{-1}\hat{X}_N^Ty = (X_N^TP_NX_N)^{-1}X_N^TP_Ny \tag{9.16}$$

Theorem 9.5.2 *The instrumental variables estimator $\hat{\theta}_{IV}$ is consistent if the conditions (9.15) are satisfied, and* $\text{var}(\hat{\theta}_{IV})$ *is approximately given by* $\sigma^2(X_N^TP_NX_N)^{-1}$.

Proof By filling in (9.4) into (9.16) it follows that

$$\hat{\theta}_{IV} = \theta + \{X_N^TZ_N(Z_N^TZ_N)^{-1}Z_N^TX_N\}^{-1}X_N^TZ_N(Z_N^TZ_N)^{-1}Z_N^T\varepsilon_N.$$

Consistency now follows immediately from the assumptions in (9.15). The expression for the variance follows from Theorem 9.2.2, replacing X by $\hat{X}_N$. □

For the ARMAX model (9.14), assuming that the input $u(t)$ only depends on the past outputs $\{y(s), s \leq t\}$, one can choose instruments from the set $\{y(s), u(s), s \leq t - r - 1\}$ as these are uncorrelated with the composite disturbance term $\varepsilon(t) + \sum_{i=1}^{r} \gamma_i\varepsilon(t-i)$. The resulting IV estimator is consistent, but it may be far from efficient.

From an asymptotic point of view, it is optimal to use maximum likelihood. Denoting the lag operator by $(z^{-1}y)(t) = y(t-1)$, the model (9.14) can be written as

$\alpha(z^{-1})y(t) = \beta(z^{-1})u(t) + \gamma(z^{-1})\varepsilon(t)$. Because the model is assumed to be invertible, $\varepsilon(t) = (\gamma(z^{-1}))^{-1}(\alpha(z^{-1})y(t) - \beta(z^{-1})u(t)) = F(y(s), u(s), s \leq t)$ for a function F that is linear in the observed data but nonlinear in the unknown parameters $\theta = (\alpha_1, \cdots, \alpha_p, \beta_0, \cdots, \beta_q, \gamma_1, \cdots, \gamma_r)$. Because $\alpha(\infty) = \gamma(\infty) = 1$, this can also be written in prediction error form

$$\varepsilon(t, \theta) = y(t) - f(\theta, y(s-1), u(s), s \leq t) \tag{9.17}$$

If the process $\varepsilon(t)$ satisfies assumption A2*, then (conditionally on starting conditions in (9.14)) the maximum likelihood estimators are obtained by minimizing $\sum_{t=m+1}^{N} \varepsilon^2(t, \theta)$ over θ, where $m = \max\{p, q, r\}$. Note that (9.17) corresponds to a nonlinear regression model of the type (9.7), so that the parameters θ can be estimated, for instance, by the Gauss-Newton iterations (9.8).

An alternative is to use the Kalman filter. For given parameter vector θ, the ARMAX system (9.14) can be expressed in state space form, see Sect. 6.6. The mean $\mu(t)$ and variance $\sigma^2(t)$ can then be computed by means of the Kalman filter, see Theorem 7.3.1 and Proposition 7.3.3. In fact, in terms of the notation of Theorem 7.3.1 and Proposition 7.3.3 we have $\mu(t) = \hat{y}(t)$ and $\sigma^2(t) = CP(t)C^T + GG^T$. Considering the inputs as fixed and using the notation $U_t = \{u(t), u(t-1), \cdots, u(1)\}$ and similarly for Y_t, the likelihood function can be written by sequential conditioning as $\log L(\theta) = \sum_{t=1}^{N} \log(p(y(t) \mid \theta, U_t, Y_{t-1})$. Under assumption A2*, the densities $p(y(t) \mid \theta, U_t, Y_{t-1})$ are normal, with mean $\mu(t) = E(y(t) \mid \theta, U_t, Y_{t-1})$ and variance $\sigma^2(t)$, so that

$$\log L(\theta) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{N}(y(t) - \mu(t))^2/\sigma^2(t) - \frac{1}{2}\sum_{t=1}^{N}\log\sigma^2(t). \tag{9.18}$$

This can then serve for a numerical optimization algorithm to obtain the maximum likelihood estimate.

The foregoing results can be generalized to multivariate systems. As mentioned in Sect. 6.3, the parameters of multivariate VARMAX systems are in general not uniquely defined. That is, there exist different parameter vectors that describe exactly the same (stochastic) input-output system. This so-called non-identifiability implies that the likelihood function is constant for such parameters, so that the gradient may be zero in such directions. This causes numerical problems, that can be solved by choosing a canonical form for the parameters. We refer to [52].

Identification methods that are based on the prediction errors as in (9.17) are called prediction error identification (PEI) methods. For multivariate systems, let $V(\theta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t, \theta)\varepsilon^T(t, \theta)$ denote the sample covariance matrix of the prediction errors. Least squares corresponds to the criterion trace$(V(\theta))$, and it can be shown that maximum likelihood corresponds to the criterion $\log(\det(V(\theta))$. So, in the case of a single output these two methods are equivalent, but for multi-output systems this only holds true if $V(\theta)$

is diagonal and there are no cross-equation parameter restrictions in the equations (9.17). The consistency and relative efficiency of PEI methods has been investigated under quite general conditions, see [47].

9.6 Model Validation

Different model specifications may lead to different estimates of the underlying system. In order to decide about the model structure, and accordingly about the estimation method to be used, we can estimate different models and perform diagnostic tests on the underlying model assumptions. In this section we discuss some of the diagnostic tools that may be helpful in this respect.

9.6.1 Lag Orders

The estimation of ARMAX models requires that the lag orders (p, q, r) in (9.14) have been specified. If the orders are chosen too large this means that many parameters have to be estimated, with a corresponding loss of efficiency. On the other hand, if the orders are too small then the estimates become inconsistent. That is, the choice of the lag orders involves a trade-off between efficiency and consistency. We illustrate this by an example.

Example 9.6.1 Consider the causal AR(2) model $y(t) = \alpha_1 y(t-1) + \alpha_2 y(t-2) + \varepsilon(t)$, where ε satisfies assumption A2. First assume that the order is specified too large, that is, that $\alpha_2 = 0$. Using the variance expression in Theorem 9.2.2, with the regressors $x(t) = (y(t-1), y(t-2))^T$, it follows that $\hat{\alpha}_1$ in the AR(2) model has variance

$$\begin{aligned}
\operatorname{var}(\hat{\alpha}_1) &= \sigma^2[(X^T X)^{-1}]_{1,1} \\
&= \frac{\sigma^2 \sum y^2(t-2)}{\sum y^2(t-1) \sum y^2(t-2) - (\sum y(t-1) y(t-2))^2} \\
&\approx \frac{\sigma^2}{N R(0)\{1 - (R(1)/R(0))^2\}} = \frac{1}{N},
\end{aligned}$$

where $R(k)$ denotes the autocovariances of the process $y(t)$. Because $\alpha_2 = 0$, there holds $R(0) = \sigma^2(1-\alpha_1^2)^{-1}$ and $R(1) = \alpha_1 R(0)$. In the correctly specified AR(1) model, the estimator has variance

$$\operatorname{var}(\hat{\alpha}_1) = \frac{\sigma^2}{\sum y^2(t-1)} \approx \frac{\sigma^2}{N R(0)} = \frac{1-\alpha_1^2}{N}.$$

This shows that too large models lead to inefficient estimators. On the other hand, if an AR(1) model is estimated while in fact $\alpha_2 \neq 0$, then

$$\text{plim}(\hat{\alpha}_1) = \text{plim}\left(\frac{\frac{1}{N}\sum y(t)y(t-1)}{\frac{1}{N}\sum y^2(t-1)}\right) = \alpha_1 + \alpha_2 \frac{R(1)}{R(0)}. \tag{9.19}$$

So in this case the estimator is inconsistent if $R(1) \neq 0$.

Several methods have been developed for choosing the lag orders. For example, if the parameters are estimated by maximum likelihood then the results in Sect. 9.3 show that the estimators are approximately normally distributed. The significance of the parameters in model (9.14) can then be evaluated by the usual t- and F-tests.

If only a single output is observed, then the order of AR(p) models and MA(q) models can be based on the (partial) autocorrelations. The autocorrelations of a stationary process are defined by $AC(k) = R(k)/R(0)$, with corresponding sample estimates $SAC(k) = \hat{R}(k)/\hat{R}(0)$. If y is an MA(q) process, then $AC(k) = 0$ for $k > q$. If y is an AR(p) process then in the regression model (9.9) of an AR(k) model there holds $\alpha_k = 0$ for $k > p$. The sample partial autocorrelations are defined by $SPAC(k) = \hat{\alpha}_k$, the parameter of $y(t-k)$ in the estimated AR(k) model for the data (including constant, trends and dummies if needed). As a rule of thumb, estimated values SAC and $SPAC$ are considered significant if they are (in absolute value) larger than $2/\sqrt{N}$, where N is the sample size.

An alternative is to use information criteria, for instance the Akaike or Bayes criterion

$$AIC = \log(\hat{\sigma}^2) + \frac{2M}{N}, \qquad BIC = \log(\hat{\sigma}^2) + \frac{M\log(N)}{N} \tag{9.20}$$

Here $\hat{\sigma}^2$ is the estimated variance of the residuals of the model, and M is the number of AR and MA parameters of the model. For instance, for a univariate ARMA(p, q) process $M = p + q$, and for the model AR(p) model (9.10) with constant and trend $M = p$. The model with the smallest value of AIC or BIC is preferred. These criteria make an explicit trade-off between bias, measured by the error variance $\hat{\sigma}^2$, and efficiency, measured by the number of parameters.

9.6.2 Residual Tests

The estimation methods in Sects. 9.4 and 9.5 are based on the assumptions A2 or A2* for the error terms. If, for example, the lag orders have been misspecified then this may result in serial correlation of the error terms. And if the data are not appropriately transformed then the error terms may show changing variance. If the error terms are not normally distributed, then least squares is no longer equivalent to maximum likelihood. In all these cases, the methods discussed in Sects. 9.4 and 9.5 may give misleading results.

Tests of these assumptions are based on the model residuals $\hat{\varepsilon}(t) = y(t) - \hat{y}(t)$, where $\hat{y}(t)$ denotes the fitted values. For instance, for the ARMAX model (9.14) $\hat{\varepsilon}(t) = y(t) - \sum_{i=1}^{p} \hat{\alpha}_i y(t-i) - \sum_{i=0}^{q} \hat{\beta}_i u(t-i) - \sum_{i=1}^{r} \hat{\gamma}_i \hat{\varepsilon}(t-i)$. It is always informative to make a time plot of the residuals to get an idea of possible misspecification. The sample autocorrelations $SAC_\varepsilon(k) = \hat{R}_\varepsilon(k)/\hat{R}_\varepsilon(0)$ give an indication of possible serial correlation, where $\hat{R}_\varepsilon(k)$ are the sample autocovariances of $\hat{\varepsilon}(t)$. As before, if there exist many values of k for which $| SAC_\varepsilon(k) | > 2/\sqrt{N}$ then this is a sign of serial correlation.

A combined test is the Box-Pierce test $Q_m = N \sum_{k=1}^{m} (SAC_\varepsilon(k))^2$. Under the null-hypothesis that the model is correctly specified, this test follows a $\chi^2_{(m-p-r)}$ distribution for large enough sample sizes. The following Ljung-Box test involves an adjustment for finite sample effects, and also follows an asymptotic $\chi^2_{(m-p-r)}$ distribution.

$$LB_m = N(N+2) \sum_{k=1}^{m} (N-k)^{-1} (SAC_\varepsilon(k))^2. \tag{9.21}$$

The null hypothesis of no serial correlation is rejected for large values of LB_m. This means that the model is not correct, and a possible solution is to enlarge the lag orders.

As concerns heteroscedasticity, it may be that the variance is related to the level of the series or that the variance shows correlation over time. Tests are based on the series of squared residuals $\hat{\varepsilon}(t)^2$. For example, if an ARX$(1, 0)$ model (9.13) is estimated then one can consider the regressions

$$\hat{\varepsilon}^2(t) = \lambda_0 + \lambda_1 y(t-1) + \lambda_2 y^2(t-1) + \lambda_3 u(t) + \lambda_4 u^2(t), \tag{9.22}$$

$$\hat{\varepsilon}^2(t) = \lambda_0 + \lambda_1 \hat{\varepsilon}^2(t-1) + \lambda_2 \hat{\varepsilon}^2(t-2). \tag{9.23}$$

These equations can of course be generalized. The null hypothesis is that $\lambda_i = 0$ for all $i \neq 0$. In both cases an F-test can be used, and under the null hypothesis the distribution is approximately $\chi^2_{(m)}$ where m is the number of restrictions ($m = 4$ in (9.22), and $m = 2$ in (9.23)). If there is significant heteroscedasticity then the data can be transformed, or one can adjust the identification criterion. More general, the following result holds true.

Theorem 9.6.1 *For the regression model* (9.3)*, assume that A1 is satisfied and that* $E(\varepsilon) = 0$ *and* $\mathrm{var}(\varepsilon) = V$ *with* V *nonsingular. Then the BLUE estimator is obtained by minimizing* $\varepsilon^T V^{-1} \varepsilon$*, with solution* $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ *and* $\mathrm{var}(\hat{\beta}) = (X^T V^{-1} X)^{-1}$.

Proof As V is a nonsingular covariance matrix, it is positive definite and has a symmetric square root $V^{\frac{1}{2}}$ such that $V^{\frac{1}{2}} V^{\frac{1}{2}} = V$. Let $y_* = V^{-\frac{1}{2}} y$, $X_* = V^{-\frac{1}{2}} X$ and $\varepsilon_* = V^{-\frac{1}{2}} \varepsilon$, then (9.3) implies that $y_* = X_* \beta + \varepsilon_*$ with $\mathrm{var}(\varepsilon_*) = I$. According to Theorem 9.2.2, the BLUE estimator is given by $\hat{\beta} = (X_*^T X_*)^{-1} X_*^T y_*$ with $\text{var}(\hat{\beta}) = (X_*^T X_*)^{-1}$, and this corresponds to the minimization of $\varepsilon_*^T \varepsilon_* = \varepsilon^T V^{-1} \varepsilon$. □

The technique to transform the data in such a way that the error term satisfies assumption A2 is called *pre-whitening*. In practice, the covariance matrix V is unknown and has to be estimated. In the case of heteroscedasticity, V is a diagonal matrix and the entries $v_{tt} = E(\varepsilon^2(t))$ can be estimated, for example, by models of the type (9.22), (9.23). The parameters β are then estimated by weighted least squares, with criterion function $\sum_{t=1}^{N} \varepsilon^2(t)/v_{tt}$.

Finally we consider the assumption of normality of the error terms. This can be tested by considering the standardized third and fourth moments of the residuals. Let $\bar{\varepsilon} = \frac{1}{N} \sum_{t=1}^{N} \hat{\varepsilon}(t)$ and $\hat{\sigma}^2 = \frac{1}{N} \sum_{t=1}^{N} (\hat{\varepsilon}(t) - \bar{\varepsilon})^2$, then $\hat{\mu}_i = \frac{1}{N} \sum_{t=1}^{N} (\hat{\varepsilon}(t) - \bar{\varepsilon})^i / \hat{\sigma}^i$ are the skewness (for $i = 3$) and kurtosis (for $i = 4$). It can be shown that, asymptotically and under the null hypothesis that A2* is satisfied, the Jarque-Bera test

$$JB = N(\frac{1}{6}\hat{\mu}_3^2 + \frac{1}{24}(\hat{\mu}_4 - 3)^2) \tag{9.24}$$

has the $\chi^2_{(2)}$ distribution. The normal distribution is symmetric (skewness zero) and has kurtosis equal to 3 (a measure of the thickness of the tails of the distribution). Normality may be rejected, for instance, because there are some excessively large residuals. They may arise because of special circumstances, for instance a measurement error or a temporary disruption of the process. Because the least squares criterion penalizes residuals by taking the squares, such outliers may have large effects on the estimates. This can be reduced by using more robust identification criteria, for example by minimizing $\sum_{t=1}^{N} | \varepsilon(t) |$.

9.6.3 Inputs and Outputs

For multivariable systems, the question arises how many equations should be estimated and what are the properties of the error process. It is usual to model either all the variables as a multivariate stochastic process or to model some of the variables (the outputs) in terms of the others (the inputs). This is also the basis for the methods described in Sects. 9.4 and 9.5. Here we will not discuss alternative modelling approaches, but we give two examples indicating the importance of these questions.

Example 9.6.2 In this example we analyse the effect of incomplete model specification. Assume that three variables are observed that actually consist of one input and two outputs, related by the equations

$$y_1(t) = \alpha_1 y_2(t) + \beta_1 y_1(t-1) + \gamma_1 u(t) + \varepsilon_1(t),$$

$$y_2(t) = \alpha_2 y_1(t) + \beta_2 y_2(t-1) + \gamma_2 u(t) + \varepsilon_2(t),$$

where $(\varepsilon_1, \varepsilon_2)^T$ is a white noise process with covariance matrix I. Suppose that we do not know that y_2 is an output and that we estimate only the first equation for y_1, seen as an ARX$(1, 0)$ model with output y_1 and inputs u and y_2. This model structure suggests to estimate the parameters by least squares, see Sect. 9.4. However, this gives inconsistent estimates. The result in Theorem 9.4.3 does not apply, because the regressor $y_2(t)$ is correlated with $\varepsilon_1(t)$ if $\alpha_2 \neq 0$. More precisely, assume that the processes y_1, y_2 and u are all stationary, and let $\theta = (\alpha_1, \beta_1, \gamma_1)^T$ and $x(t) = (y_2(t), y_1(t-1), u(t))^T$. Then the least squares estimator $\hat{\theta}_N$ in the equation for y_1 has the property that $\text{plim}(\hat{\theta}_N) = \theta + V^{-1}\delta$, where $V = \text{var}(x(t))$ is invertible and $\delta \in \mathbb{R}^3$ has as first entry $E(y_2(t)\varepsilon_1(t))$. Taking into account the two model equations, it follows that $E(y_2(t)\varepsilon_1(t)) = \alpha_2/(1-\alpha_1\alpha_2) \neq 0$. This is called the simultaneity bias, that arises when some of the system equations are missing in the model.

Example 9.6.3 Next we analyse the consequences of a wrong specification of the properties of the error process. Suppose that the system consists of a single input and a single output that are both measured with error, for instance,

$$y(t) = y_*(t) + \varepsilon_1(t), \quad u(t) = u_*(t) + \varepsilon_2(t), \quad y_*(t) = \beta u_*(t-1) + \varepsilon_3(t).$$

Here the underlying system for the unobserved variables (y_*, u_*) is ARX$(0, 1)$. We assume that ε_i are independent white noise processes with zero mean and variance σ_i^2, $i = 1, 2, 3$, and that u_* is a stationary process with mean zero and variance σ_*^2 that is independent of ε_i, $i = 1, 2, 3$. In terms of the observed input and output, the ARX$(0, 1)$ model $y(t) = \theta u(t-1) + \varepsilon(t)$ is correctly specified, in so far as the lag order is correct, the input and output are chosen correctly, and the errors satisfy assumption A2. Indeed, actually $y(t) = \beta u(t-1) + \varepsilon(t)$ where $\varepsilon(t) = \varepsilon_1(t) - \beta\varepsilon_2(t-1) + \varepsilon_3(t)$ is a white noise process. However, the least squares estimator is not consistent because the orthogonality condition of Theorem 9.2.3 is not satisfied. As $E(\varepsilon(t)u(t-1)) = -\beta\sigma_2^2$ and $E(u^2(t-1)) = \sigma_*^2 + \sigma_2^2$, it follows that

$$\text{plim}(\hat{\theta}_N) = \beta - \frac{\beta\sigma_2^2}{\sigma_*^2 + \sigma_2^2} = \beta(1 - \frac{1}{S+1}),$$

where $S = \sigma_*^2/\sigma_2^2$ is the so-called signal-to-noise ratio for the input. This shows that a wrong specification of the error assumptions may lead to inconsistent results. Especially when the noise is relatively large, that is, when S is small, the estimates may be very unreliable. Note that the orthogonality condition can not be checked by computing the correlation between the regressor $u(t-1)$ and the residuals $\hat{\varepsilon}(t) = y(t) - \hat{\theta}u(t-1)$, because $\text{plim}(\frac{1}{N}\sum_{t=1}^{N}\hat{\varepsilon}(t)u(t-1)) = E(y(t)u(t-1)) - \text{plim}(\hat{\theta}_N)E(u^2(t-1)) = 0$.

9.6.4 Model Selection

In system identification one is confronted with the choice of data, model class, estimation method, and tools for evaluating the model quality. The validation techniques for the lag orders and the residuals discussed in Sects. 9.6.1 and 9.6.2 are of help. Further, the intended model use may suggest additional evaluation criteria. For instance, if forecasting is the objective then the models can be compared with respect to their forecast performance. The standard deviation

$$\hat{\sigma} = \{\frac{1}{N}\sum_{t=1}^{N}\hat{\varepsilon}(t)^2\}^{\frac{1}{2}} \tag{9.25}$$

is an indication of this. However, in least squares the data are first used to minimize $\hat{\sigma}$, so that this may underestimate the future forecast errors. A more reliable criterion is $\sigma^* = \{\frac{1}{N}\sum_{t=1}^{N}\varepsilon^*(t)^2\}^{\frac{1}{2}}$, where $\varepsilon^*(t) = y(t) - y^*(t)$ is the residue corresponding to the model that is estimated using the data $\{y(s-1), u(s), s \leq t\}$. The disadvantage is that this requires the estimation of a sequence of models, a new one for every additional observation. One can also consider m-step-ahead prediction, where only the data $\{y(s-1), u(s), s \leq t-m\}$ are used to estimate a model to forecast $y(t)$. Instead of quadratic criteria one can also consider the absolute errors $\frac{1}{N}\sum_{t=1}^{N} \mid \varepsilon(t) \mid$ or the relative errors $\frac{1}{N}\sum_{t=1}^{N}(\mid \varepsilon(t) \mid / \mid y(t) \mid)$. For input-output systems that allow experiments with the inputs, one can also compare the simulated outputs of the model with the outputs that result in reality.