criterion. Although it may be difficult to come up with an exact description of the plant in reality, studying these methods can provide some useful insights into the performance of empirical methods like the prediction error minimization. We present the two most popular methods here.

7.2.3.1 Maximum Likelihood Estimation

In system identification, one is trying to extract system information out of measurements that are inherently unreliable. In maximum likelihood estimation, this is formalized by describing each observation as a realizaton of a random variable with certain probability distribution. For instance, if we assume a model

$$y(k) = \phi^T(k)\theta + \varepsilon(k) \tag{7.49}$$

where $\varepsilon(k)$ is a Gaussian variable with zero mean and variance r_ε, then the probability density function (PDF) of $y(k)$ becomes

$$dF(\zeta; y(k)) = \frac{1}{\sqrt{2\pi r_\varepsilon}} \exp\left\{-\frac{(\zeta - \phi^T(k)\theta)^2}{2r_\varepsilon}\right\} \tag{7.50}$$

In the above, ζ represents a particular realized value for $y(k)$.

In parametric identification with N data points, we can work with a joint PDF for $Y_N \triangleq (y(1), \cdots, y(N))$. Let us denote the joint PDF as $dF(\zeta_N; Y_N)$. Again, ζ^N is a variable representing realization of Y_N. Suppose the actual observations are given as $\hat{Y}_N = (\hat{y}(1), \cdots, \hat{y}(N))$. Once we insert these values into the probability density function, $dF(\hat{Y}_N; Y_N)$ is now a deterministic function of θ called "likelihood function." We denote the likelihood function for the observation $\hat{Y}_N$ as $\ell(\theta|\hat{Y}_N)$.

The basic idea of maximum likelihood estimation is to make the observations "as likely as possible" by choosing θ such that the liklihood

function is maximized. In other words,

$$\hat{\theta}_N^{ML} = \arg\left\{\max_\theta \ell(\theta|\hat{Y}_N)\right\} \tag{7.51}$$

Often, it is generally quite difficult to derive the likelihood function from a stochastic system model. An exception is the case when the model can be put into a linear predictor form in which the observation is linear with respect to both the unknown parameters and random variables.

Let us apply the maximum likelihood method to the following linear identification problem:

$$Y_N = \Phi_N\theta + \mathcal{E}_N \tag{7.52}$$

In the above, we assume that $\mathcal{E}_N$ is a zero-mean Gaussian variable vector of covariance $R_\mathcal{E}$. Then, we have

$$\begin{aligned} dF(\hat{Y}_N; Y_N) &= dF(\hat{Y}_N - \Phi_N\theta; \mathcal{E}_N) && (7.53)\\ &= \frac{1}{\sqrt{(2\pi)^N \det(R_\mathcal{E})}} \exp\left\{-\frac{1}{2}(\hat{Y}_N - \Phi_N\theta)^T R_\mathcal{E}^{-1}(\hat{Y}_N - \Phi_N\theta)\right\} \end{aligned}$$

The maximum likelihood estimator is defined as

$$\begin{aligned} \hat{\theta}_N^{ML} &= \arg\left\{\max_\theta dF(\hat{Y}_N; Y_N)\right\} && (7.54)\\ &= \arg\left\{\max_\theta \log\left(dF(\hat{Y}_N; Y_N)\right)\right\} && (7.55)\\ &= \arg\left\{\max_\theta \left(-\frac{1}{2}(\hat{Y}_N - \Phi_N\theta)^T R_\mathcal{E}(\hat{Y}_N - \Phi_N\theta)\right)\right\} && (7.56)\\ &= \arg\left\{\min_\theta \left(\frac{1}{2}(\hat{Y}_N - \Phi_N\theta)^T R_\mathcal{E}(\hat{Y}_N - \Phi_N\theta)\right)\right\} && (7.57) \end{aligned}$$

Note that the above is a weighted least squares estimator. We see that, when the weighting matrix is chosen as the inverse of the covariance matrix for the output error term $\mathcal{E}_N$, the weighted least squares estimation is equivalent to the maximum likelihood estimation. In addition, the unweighted least squares estimator is a maximum likelihood estimator for

the case when the output error is an i.i.d. Gaussian sequence (in which case the covariance matrix for $\mathcal{E}_N$ is in the form of $r_\varepsilon I_N$).

7.2.3.2 Bayesian Estimation

Bayesian estimation is a philosophically different approach to the parameter estimation problem. In this approach, parameters themselves are viewed as random variables with a certain prior probability distribution. If the observations are described in terms of the parameter vector, the probability distribution of the parameter vector changes after the observations. The distribution after the observations is called posterior probability distribution, which is given by the conditional distribution of the parameter vector conditioned with the observation vector. The parameter value for which the posterior PDF attains its maximum is called the "maximum a posteriori (MAP) estimate." It is also possible to use the mean of the posterior distribution as an estimate, which gives the "minimum variance estimate."

One of the useful rules in computing the posterior PDF is Bayes's rule. Let us denote the conditional PDF of the parameter vector for given observations as $dF(\hat{\theta}|\zeta_N;\theta|Y_N)$. Then, Bayes's rule says

$$dF(\hat{\theta}|\zeta_N;\theta|Y_N) = \frac{dF(\zeta_N|\hat{\theta};Y_N|\theta)\cdot dF(\hat{\theta};\theta)}{dF(\zeta_N;Y_N)} \tag{7.58}$$

$dF(\zeta_N;Y_N)$ is independent of θ and therefore is constant once it is evaluated for given observation $\hat{Y}_N$. Hence, the MAP estimator becomes

$$\hat{\theta}_N^{MAP} = \arg\left\{\max_{\hat{\theta}} dF(\zeta_N|\hat{\theta};Y_N|\theta)\cdot dF(\hat{\theta};\theta)\right\} \tag{7.59}$$

Note that we end up with a parameter value that maximizes the product of the likelihood function and the prior density.

Let us again apply this concept to the linear parameter estimation problem of

$$Y_N = \Phi_N \theta + \mathcal{E}_N \tag{7.60}$$

where $\mathcal{E}_N$ is a Gaussian vector of zero mean and covariance $R_\mathcal{E}$. We also treat θ as a Gaussian vector of mean $\hat{\theta}(0)$ and covariance $P(0)$. Hence, the prior distribution is a normal distribution of the above mean and covariance.

Next, let us evaluate the posterior PDF using Bayes's rule.

$$dF(\hat{\theta}|\hat{Y}_N;\theta|Y_N) = [\text{constant}] \times dF_\mathcal{N}(\hat{Y};Y_N)_{(\Phi_N\theta,R_\mathcal{E})} \cdot dF_\mathcal{N}(\hat{\theta},\theta)_{(\hat{\theta}(0),P(0))} \tag{7.61}$$

where

$$dF_\mathcal{N}(\hat{x},x)_{(\bar{x},R)} = \frac{1}{\sqrt{(2\pi)^N \det(R)}} \exp\left\{-\frac{1}{2}(\hat{x}-\bar{x})^T R^{-1}(\hat{x}-\bar{x})\right\} \tag{7.62}$$

The MAP estimate can be obtained by maximizing the logarithm of the posterior PDF:

$$\begin{aligned}\hat{\theta}_N^{MAP} &= \arg\left\{\max_{\hat{\theta}}\left(-\frac{1}{2}(\hat{Y}_N-\Phi_N\hat{\theta})^T R_\mathcal{E}^{-1}(\hat{Y}_N-\Phi_N\hat{\theta}) - \frac{1}{2}(\hat{\theta}-\hat{\theta}(0))^T P^{-1}(0)(\hat{\theta}-\hat{\theta}(0)\right)\right\} \\ &= \arg\left\{\min_{\hat{\theta}}\frac{1}{2}\left((\hat{Y}_N-\Phi_N\hat{\theta})^T R_\mathcal{E}^{-1}(\hat{Y}_N-\Phi_N\hat{\theta})\right) + (\hat{\theta}-\hat{\theta}(0))^T P^{-1}(0)(\hat{\theta}-\hat{\theta}(0)\right\}\end{aligned} \tag{7.63}$$

Solving the above least squares problem, we obtain

$$\hat{\theta}_N^{MAP} = \left(\Phi_N^T R_\mathcal{E}^{-1}\Phi_N + P^{-1}(0)\right)^{-1}\left(\Phi_N^T R_\mathcal{E}^{-1}\hat{Y}_N + P^{-1}(0)\hat{\theta}(0)\right) \tag{7.64}$$

Using the Matrix Inversion Lemma, one can rewrite the above as

$$\hat{\theta}_N^{MAP} = \hat{\theta}(0) + P(0)\Phi_N^T\left(\Phi_N^T P(0)\Phi_N + R_\mathcal{E}\right)^{-1}\left(\hat{Y}_N - \Phi_N\hat{\theta}(0)\right) \tag{7.65}$$

We make the following observations:

- The above indicates that, as long as $P(0)$ is chosen as a nonsingular

matrix and the persistent excitation condition is satisfied, $\hat{\theta}_N^{MAP}$ converges to $\hat{\theta}_N^{LS}$ as $N \to \infty$. Hence, all the asymptotic properties of the least squares identification apply to the above method as well.

- If $P(0)$ is chosen as a singular matrix, the estimate of θ may be *biased* since the null space of $P(0)$ represents the parameter subspace corresponding to zero update gain.
- From (7.63), we see that specifying the initial parameter covariance matrix $P(0)$ to be other than ∞I is equivalent to penalizing the deviation from the initial parameter guess through weighting matrix $P^{-1}(0)$ in the least squares framework. The standard least squares solution is interpreted in the Bayesian framework as the MAP solution corresponding to a uniform initial parameter distribution (i.e., $P(0) = \infty I$).

Utilizing prior knowledge in the above framework can help us obtain a smoother and more realistic impulse response. In Section **??**, we suggested using a diagonal weighting matrix to penalize the magnitudes of the impulse response coefficients so that a smoother step response can be obtained. We now see that this is equivalent to specifying the initial parameter covariance as a diagonal matrix (*i.e.*, the inverse of the weighting matrix) in the Bayesian framework. The statistical interpretation provides a formal justification for this practice and a systematic way to choose the weighting matrix (possibly as a nondiagonal matrix).

(7.65) can be written in the following recursive form:

$$\begin{aligned} \hat{\theta}(k) &= \hat{\theta}(k-1) + K(k)\left(y(k) - \phi^T(k)\hat{\theta}(k-1)\right) \\ K(k) &= \frac{P(k-1)\phi(k)}{1+\phi^T(k)P(k-1)\phi(k)} \\ P(k) &= P(k-1) - \frac{P(k-1)\phi(k)\phi^T(k)P(k-1)}{1+\phi^T(k)P(k-1)\phi(k)} \end{aligned} \tag{7.66}$$

where $\hat{\theta}(k)$ represents $\hat{\theta}_k^{MAP}$ or $E\{\theta|Y_k\}$ and

$P(k) = E\left\{(\theta - \hat{\theta}(k))(\theta - \hat{\theta}(k))^T | Y_k\right\}$. The above formula is easily derived by formulating the problem as a special case of state estimation and applying the Kalman filtering.

One could generalize the above to the time-varying parameters by using the following system model for parameter variation:

$$\begin{aligned} \theta(k) &= \theta(k-1) + w(k) \\ y(k) &= \phi^T(k)\theta(k) + \nu(k) \end{aligned} \tag{7.67}$$

where $w(k)$ is white noise. This way, the parameter vector $\theta(k)$ can be assumed to be time-varying in a random walk fashion. One may also model $w(k)$ and $\nu(k)$ as nonwhite signals by further augmenting the state vector as described earlier

We will demonstrate an application of the Bayesian approach to the impulse response coefficient identification through the following example.

Example:
In practice, it may be more appropriate to assume (in prior to the identification) the derivatives of the impulse response as zero-mean random variables of Gaussian distribution and specify the covariance of the derivative of the impulse response coefficients. In other words, one may specify

$$E\left\{\frac{dh}{dt}_{t=i\cdot T_s}\right\} \approx E\left\{\frac{h_i - h_{i-1}}{T_s}\right\} = 0; \qquad 1 \le i \le n \tag{7.68}$$

$$E\left\{\left(\frac{dh}{dt}_{t=i\cdot T_s}\right)^2\right\} \approx E\left\{\left(\frac{h_i - h_{i-1}}{T_s}\right)^2\right\} = \frac{\sigma_i}{T_s^2} \tag{7.69}$$

In this case, $P(t_0)$ (the covariance for θ) takes the following form:

$$P(t_0) = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_n \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & -1 & 1 \end{bmatrix}^T \right)^{-1} \tag{7.70}$$

Note that the above is translated as penalizing the 2-norm of the difference between two successive impulse response coefficients in the least squares identification method. It is straightforward to extend the above concepts and model the second order derivatives of the impulse response as normally distributed zero-mean random variables.

(Comment: ADD NUMERICAL EXAMPLE HERE!!!)

7.2.4 OTHER METHODS

There are other methods for estimating parameters in the literature. Among them, a method that stands out is the *instrumental variable (IV) method*. The basic idea behind this method is that, in order for a model to be good, the prediction error must show little or no correlation with past data. If they show significant correlation, it implies that there is information left over in the past data not utilized by the predictor.

In the IV method, a set of variables called "instruments" (denoted by vector η hereafter) must be defined first. η contains some transformations of past data $(y(k-1), \cdots, y(0), u(k-1), \cdots, u(0))$. Then, θ is determined from the following relation:

$$\frac{1}{N} \sum_{k=1}^{N} \eta(k) e_{pred}(k, \theta) = 0 \tag{7.71}$$

$\eta(k)$ is typically chosen to be of same dimension as the parameter vector θ. This way, one obtains the same number of equations as unknowns. Sometimes, η is chosen to be of higher dimension. Then, θ can be determined by minimizing some norm of $\frac{1}{N}\sum_{k=1}^{N}\eta(k)e_{pred}(k,\theta)$. Filtered e_{pred} can be used as well in the above. The success of the method obviously depends on the choice of instruments. See Ljung (1987) for guidelines on how to choose them. If $\eta(k)$ is chosen as $\phi(k)$, one obtains the same estimate as the least squares estimate. It is also possible to choose η that contains parameters. This leads to pseudo-linear regression.

Other variations to the least squares regression is the so called *biased regression* methods in which the regression is restricted to a subspace of the parameter space. The subspace is not chosen *a priori*, but is formed by incrementally adding on a one-dimensional space chosen to maximize the covariance of data ϕ (as in the *Principal Component Regression*) or to maximize the covariance between ϕ and y (as in the *Partial Least Squares*). These methods are designed to reduce the variance (esp. when the data do not show adequate excitation of the whole parameter space) at the expense of bias. In the Bayesian estimation setting, this can be interpreted as choosing a singular initial covariance matrix $P(0)$. However, the singular directions are determined on the basis of data rather than prior knowledge.

7.3 NONPARAMETRIC IDENTIFICATION METHODS

When one has little prior knowledge about the system, nonparametric identification which assumes very little about the underlying system is an alternative. Nonparametric model structures include frequency response models, impulse response models, *etc.*. These model structures intrinsically have no finite-dimensional parameter representations. In reality, however,

the dividing line between the parametric identification and the nonparametric identification is somewhat blurred: In nonparametric identification, some assumptions are always made about the system structure (*e.g.*, a finite length impulse response, smoothness of the frequency response) to obtain a well-posed estimation problem. In addition, in parametric identification, a proper choice of model order is often determined by examining the residuals from fitting models of various orders.

7.3.1 FREQUENCY RESPONSE IDENTIFICATION

Dynamics of a general linear system can be represented by the system's frequency response, which is defined through amplitude ratio and phase angle at each frequency. The frequency response information is conveniently represented as a complex function of ω whose modulus and argument define the amplitude ratio and the phase angle respectively. Such a function can be easily derived from the systems transfer function $G(q)$ by replacing q with $e^{j\omega}$. Hence, the amplitude ratio and phase angle of the system at each frequency is related to the transfer function parameters through the following relations:

$$A.R.(\omega) = |G(e^{j\omega})| = \sqrt{\mathrm{Re}\{G(e^{j\omega})\}^2 + \mathrm{Im}\{G(e^{j\omega})\}^2} \tag{7.72}$$

$$P.A.(\omega) = |G(e^{j\omega})| = \tan^{-1}\left[\frac{\mathrm{Im}\{G(e^{j\omega})\}}{\mathrm{Re}\{G(e^{j\omega})\}}\right] \tag{7.73}$$

Since $G(e^{j\omega})$ $(0 \leq \omega \leq \pi$ for system with sample time of 1) defines system dynamics completely, one approach to system identification is to identify $G(e^{j\omega})$ directly. This belongs to the category of nonparametric identification as frequency response is not parametrized by a finite-dimensional parameter vector (there are infinite number of frequency points).

7.3.1.1 Frequency Response Computation

The most immediate way to identify the frequency response is through a sine-wave testing, where sinusoidal perturbations are made directly to system input at different frequencies. Although conceptually straightforward, this method is of limited value in practice since (1) sinusoidal perturbations are difficult to make in practice, and (2) each experiment gives frequency response at only a single frequency.

A more practical approach is to use the results from the Fourier analysis. From the z-domain input / output relationship, it is immediate that, for system $y(k) = G(q)u(k)$,

$$G(e^{j\omega}) = \frac{Y(\omega)}{U(\omega)} \tag{7.74}$$

where

$$Y(\omega) = \sum_{k=1}^{\infty} y(k)e^{-j\omega k} \tag{7.75}$$

$$U(\omega) = \sum_{k=1}^{\infty} u(k)e^{-j\omega k} \tag{7.76}$$

Hence, by dividing the Fourier transform of the output data with that of the input data one can compute the system's frequency response. What complicates the frequency response identification in practice is that one only has finite length data. In addition, output data are corrupted by noise and disturbances.

Let us assume that the underlying system is represented by

$$y(k) = G(q)u(k) + e(k) \tag{7.77}$$

where $e(k)$ is a zero-mean stationary sequence and collectively describes the

effect of noise and disturbance. We define

$$Y_N(\omega) \triangleq \frac{1}{\sqrt{N}} \sum_{k=1}^{N} y(k) e^{-j\omega k} \tag{7.78}$$

$$U_N(\omega) \triangleq \frac{1}{\sqrt{N}} \sum_{k=1}^{N} u(k) e^{-j\omega k} \tag{7.79}$$

Then,

$$G_N(\omega) \triangleq \frac{Y_N(\omega)}{U_N(\omega)} = G(e^{j\omega}) + \frac{R_N(\omega)}{U_N(\omega)} + \frac{E_N(\omega)}{U_N(\omega)} \tag{7.80}$$

where $|R_N(\omega)| = \frac{c_1}{\sqrt{N}}$ for some c_1 (Ljung, 1987). $G_N(\omega)$ computed as above using N data points is an estimate of the true system frequency response $G(e^{j\omega})$ and will be referred to as the "Empirical Transfer Function Estimate (ETFE)."

7.3.1.2 Statistical Properties of the ETFE

Let us take expectation of (7.80):

$$E\{G_N(\omega)\} = E\left\{G(e^{j\omega}) + \frac{R_N(\omega)}{U_N(\omega)} + \frac{E_N(\omega)}{U_N(\omega)}\right\} = G(e^{j\omega}) + \frac{R_N(\omega)}{U_N(\omega)} \tag{7.81}$$

We can also compute the variance as

$$E\left\{\left(G_N(\omega) - G(e^{j\omega})\right)\left(G_N(\omega) - G(e^{-j\omega})\right)\right\} = \frac{\Phi_e + \rho_N}{|U_N(\omega)|^2} \tag{7.82}$$

where $\rho_N \leq \frac{c_2}{N}$ (Ljung, 1987).

The implications of the above are as follows:

- Since the second term of the RHS of (7.81) decays as $\frac{1}{\sqrt{N}}$, $G_N(\omega)$ is an *asymptotically unbiased estimate* of $G(e^{j\omega})$.
- If $u(k)$ is a periodic signal with period of N, $|U_N(\omega)|$ is nonzero only at

N frequency points (at $\omega = \frac{2\pi \cdot k}{N}, k = 0, \cdots, N-1$). This means that the ETFE is defined only at the N frequency points. $|U_N(\omega)|$ at these frequency points keeps growing larger as $N \to \infty$, and from (7.82), we see that the variance goes to zero.

- If $u(k)$ is a randomly generated signal, as N increases,the number of frequency points at which the ETFE can be computed also increases. However, $|U_N(\omega)|^2$ is a function that fluctuates around the spectrum of $u(k)$ and therefore does not increase with data. From (7.82), we conclude that the variance does not decay to zero. This is characteristic of any nonparameteric identification where, roughly speaking, one is trying to estimate infinite number of parameters.

A practical implication of the last comment is that the estimate can be very sensitive to noise in the data (no matter how many data points are used). Hence, some smoothing is needed. The following are some simple smoothing methods:

- Select a finite number of frequency points, $\omega_1, \cdots \omega_N$ between 0 and π. Assume that $G(e^{j\omega})$ is constant over $\omega_i - \delta\omega \leq \omega \leq \omega_i + \delta\omega$. Hence, the EFTE ($G_N(\omega)$) obtained within this window are averaged, for instance, according to the signal-to-noise ratio $\frac{\Phi_e}{|U_N(\omega)|^2}$. Since the number of frequency response parameters become finite under the assumption, the variance decays to zero as $1/N$. However, the assumption leads to bias.

- A generalization of the above is to use the weighting function $W_s(\zeta - \omega)$ for smoothing. The ETFE is smoothed according to

$$G_N^s(\omega) = \frac{\int_{-\pi}^{\pi} W_s(\zeta - \omega) G_N(\omega) \frac{|U_N(\zeta)|^2}{\Phi_e(\omega)} d\zeta}{\int_{-\pi}^{\pi} W_s(\zeta - \omega) \frac{|U_N(\zeta)|^2}{\Phi_e(\omega)} d\zeta} \tag{7.83}$$

  W_s is a function that is centered around zero and is symmetric. It usually includes a parameter that determines the width of the

smoothing window and therefore the trade-off between bias and variance. Larger window reduces variance, but increases bias and vice versa. For typical choices of W_s, see Table 6.1 of Ljung (1987). Again, the variance can be shown to decay as $1/N$ under a nonzero smoothing window.

7.3.2 IMPULSE RESPONSE IDENTIFICATION

Impulse response identification is another form of nonparametric identification, that is commonly used in practice. Suppose the underlying system is described by convolution model

$$y(k) = \sum_{i=1}^{\infty} H_i u(k-i) + e_k \tag{7.84}$$

Now post-multiply $u^T(k-\tau)$ to the above equation to obtain

$$y(k)u^T(k-\tau) = \sum_{i=1}^{\infty} H_i u(k-i)u^T(k-\tau) + e(k)u^T(k-\tau) \tag{7.85}$$

Summing up the data from $k=1$ to $k=N$,

$$\left(\frac{1}{N}\sum_{k=1}^{N} y(k)u^T(k-\tau)\right) = \sum_{i=1}^{\infty} H_i \left(\frac{1}{N}\sum_{k=1}^{N} u(k-i)u^T(k-\tau)\right) + \left(\frac{1}{N}\sum_{k=1}^{N} e(k)u^T(k-\tau)\right) \tag{7.86}$$

Assuming the input had remained at the steady-state value (i.e., $u(k)=0$ for $k \le 0$), the above can be represented by

$$R_{yu}(\tau) = \sum_{i=1}^{\infty} H_i R_{uu}(\tau - i) + R_{eu}(\tau) \tag{7.87}$$

where

$$R_{yu}(\tau) = \frac{1}{N}\sum_{k=1}^{N} y(k)u^T(k-\tau) \tag{7.88}$$

$$R_{uu}(\tau) = \frac{1}{N}\sum_{k=1}^{N} u(k)u^T(k-\tau) \tag{7.89}$$

$$R_{eu}(\tau) = \frac{1}{N}\sum_{k=1}^{N} e(k)u^T(k-\tau) \tag{7.90}$$

The above equation can also be derived from a statistical argument. More specifically, we can take expectation of (7.87) to obtain

$$E\{y(k)u^T(k-\tau)\} = \sum_{i=1}^{\infty} H_i E\{u(k-i)u^T(k-\tau)\} + E\{e(k)u^T(k-\tau)\} \tag{7.91}$$

Assuming $\{u(k)\}$ and $\{e(k)\}$ are stationary sequences, R_{uu}, R_{yu} and R_{eu} are estimates of the expectations based on N data points.

Now, let us assume that $\{u(k)\}$ is a zero-mean stationary sequence that is uncorrelated with $\{e(k)\}$, which is also stationary (or $\{e(k)\}$ is a zero-mean stationary sequence uncorrelated with $\{u(k)\}$). Then, $R_{eu}(\tau) \to 0$ as $N \to \infty$. Let us also assume that $H_i = 0$ for $i > n$. An appropriate choice of n can be determined by examining $R_{yu}(\tau)$ under a white noise perturbation. When the input perturbation signal is white, $R_{uu}(i) = 0$ except $i = 0$. From the above, it is clear that $R_{yu}(\tau) =)$ if $H_\tau = 0$. Hence, one can choose n where $R_{yu} \approx 0$ for $\tau > n$.

With these assumptions, as $N \to \infty$, we can write (7.87) as

$$\begin{bmatrix} R_{yu}(1) & R_{yu}(2) & \cdots & R_{yu}(n) \end{bmatrix} \tag{7.92}$$

$$\approx \begin{bmatrix} H_1 & H_2 & \cdots & H_n \end{bmatrix} \begin{bmatrix} R_{uu}(0) & R_{uu}(1) & \cdots & R_{uu}(n-1) \\ R_{uu}(-1) & R_{uu}(0) & \cdots & R_{uu}(n-1) \\ \vdots & \ddots & \ddots & \vdots \\ R_{uu}(-n+1) & R_{uu}(-n+2) & \cdots & R_{uu}(0) \end{bmatrix}$$

Taking transpose of the above equation and rearranging it gives

$$\begin{bmatrix} H_1^T \\ H_2^T \\ \vdots \\ H_n^T \end{bmatrix} \approx \begin{bmatrix} R_{uu}(0) & R_{uu}(1) & \cdots & R_{uu}(n-1) \\ R_{uu}(-1) & R_{uu}(0) & \cdots & R_{uu}(n-2) \\ \vdots & \ddots & \ddots & \vdots \\ R_{uu}(-n+1) & R_{uu}(-n+2) & \cdots & R_{uu}(0) \end{bmatrix}^{-1} \begin{bmatrix} R_{yu}^T(1) \\ R_{yu}^T(2) \\ \vdots \\ R_{yu}^T(n) \end{bmatrix} \tag{7.93}$$

With finite-length data, parameter variance can be significant. However, because we limited the number of impulse response coefficients to n by assuming $H_i = 0, i > n$, the variance decays as $1/N$ (assuming the matrix Φ remains nonsingular). However, some bias results because of the truncation. Again, the choice of n determines the trade-off between the variance and the bias.

Note that (7.93) gives the same estimate as the least squares identification. In the case that $\{e(k)\}$ is nonstationary due to integrating type disturbances, differenced data, $\Delta y(k)$ and $\Delta u(k)$, can be used as before.

7.3.3 SUBSPACE IDENTIFICATION

There applications where it is necessary to embed into the model disturbance correlations among different outputs. In this case, MIMO identification (rather than SISO or MISO identification) is needed. Transfer function models are difficult to work with in this context, since it gives rise to a numerically ill-conditioned, nonlinear estimation problem with possible local minima. In addition, significant prior knowledge (*e.g.*, the system order, the observability index) is needed to obtain a model parameterization. An alternative is to identify a state-space model directly, using a *subspace identification* method. Different subspace identification algorithms available in the literature share the same basic concept, which