CHAPTER 4

SELECTION OF A METHODOLOGY APPLICABLE TO MPC CONTROLLED PLANTS

4.1 INTRODUCTION

In this chapter the most appropriate SID methodology is chosen for the identification of plants, controlled by MPC controllers, from measured closed-loop data.

The merits and weakness of the different methods and options regarding closed-loop system identification are considered. A methodology is selected by taking into account the characteristics of MPC controllers and industrial plants. The selection of an identification approach is discussed in Section 4.2 and the way in which identifiability can be guaranteed is discussed in Section 4.3.

One of the options for the guarantee of identifiability is the novel inter-sampling approach. An evaluation, through a simulation study, of the model variance is performed. The aim is to determine the value of this new approach. Therefore the results and conclusions from this study are presented in Section 4.4.

In Section 4.5 the selection of a type of model structure is discussed and the selected validation method is discussed in Section 4.6.

For the sake of completeness this discussion is ended with Section 4.7 in which a motivation is given for the use of the SID toolbox instead of custom-written algorithms or extra software, in the implementation of the selected identification methodology. The selected closed-loop SID methodology is then summarised in a step by step description of the method in Section 4.8.

4.2 CLOSED-LOOP IDENTIFICATION APPROACH

Common assumptions in closed-loop identification are that the existing controller is linear and the process is single variable. Practical MPC controllers are usually nonlinear because, invariably, plant inputs and outputs are constrained. Plants are often also multivariable, making many closed-loop identification results not suitable for MPC applications [3]. This is also the case for the MIBK plant, since there are constraints on the plant inputs and outputs, making the controller nonlinear, and the plant is, of course, also multivariable.

As stated in Section 3.6, the indirect and the joint input-output approaches are typically only used when the feedback law is linear. Lakshminarayanan *et al.* [40] emphasize this fact by stating that irrespective of the identification algorithm, consistent estimation from the joint input-output method is obtainable only when the controller is linear and time-invariant. Although a modification of these methods can sometimes be applied when the regulator is nonlinear, this will cause considerably more work. The MPC controller is also very complex compared to other controllers, since the MPC algorithm computes the control action at each sampling interval using on-line optimisation [28]. This fact makes the indirect and the joint input-output approaches even less attractive for MPC.

There is one variant of the joint input-output approach that does work with nonlinear controllers. This is the projection method, devised by Forssell and Ljung [14]. However, this method depends on the derivation of noncasual FIR models, which are nonparametric. According to Zhu *et al.* [41], parametric models are better for use in industrial process identification, since these models are more accurate and precise, require shorter test time, and can be more user friendly than nonparametric models. Therefore, for industrial processes, which are the processes usually controlled by MPC controllers, nonparametric models are not ideal, which, in turn, makes the projection method not ideal.

According to Ljung [11], the direct approach should be seen as the natural approach to closed-loop data analysis. The main reasons are:

- this method works regardless of the complexity of the controller,
- no special algorithms and software are required,
- consistency and optimal accuracy are obtained if the model structure contains the true system, including the noise properties, and
- unstable systems can be handled without problems, as long the closed-loop system is stable and the predictor is stable.

For a stable predictor any unstable poles of the plant model $G(q)$ must be shared by the noise model $H(q)$. Parametric models like ARX and ARMAX satisfy this constraint [11].

The only drawback with the direct approach is that a good noise model is needed in order to prevent a bias in the estimate $\hat{G}_N$ [6, 17]. Thus, as stated by Lakshminarayanan *et al.* [40], if the process model and disturbance model structures are completely known, then direct identification is the obvious choice as it can provide model parameters with the least variance and with considerably less effort.

In the case of closed-loop SID of plants controlled by MPC controllers, the identification will usually be re-identification, since the first identification takes place before the MPC controller is designed and implemented. This, in turn, implies that the model and disturbance model structures are known when closed-loop SID is performed. Therefore, the direct closed-loop identification method is the best choice when MPC controllers are involved.

When the available data sets for the MIBK plant are considered then the direct method is actually the only choice. The reason being that there are no measured reference signals available and not enough information to model the controller, as required to employ the indirect, joint input-output and projection method.

4.3 Guarantee of Identifiability

The purpose of feedback is to make the sensitivity function $So(q)$ small, especially at frequencies where the disturbance signals $v(t)$ have energy. As stated in Section 3.7.7, feedback will worsen the measured data's information about the system at these frequencies [11].

However, for the direct approach an experiment will still be informative in each of the following situations [31]:

- the reference signal is persistently exciting,
- the controller is of sufficiently high order, or
- the controller switches between several settings during the experiment.

Although a multivariable MPC changes its structure as it deals with changes in the active constraints, this cannot guarantee a given number of changes in the controller settings, since these changes are unpredictable under normal operation. Making deliberate changes to a multivariable MPC may have unforeseen consequences for process operation, and it does not provide the same control over the SNR as adding external test signals [10]. Therefore, the best way to satisfy the identifiability condition is to ensure that the reference signal is persistently exciting. For plants controlled by MPC controllers, this has been done in many different ways:

Doma *et al.* [10] satisfy this condition by adding as many different statistically independent external test signals as there are independent variables to the reference signals. Since simultaneous application of the test signals is, according to them, undesirable for complex multivariable processes, because of unforeseen interaction with unknown multivariable disturbances, they add the external tests signals one variable at a time.

However, data from a single variable test may not contain sufficient information about the multivariable character of the process, i.e. ratio between the different variables. Multivariable testing can solve this problem [3]. Therefore, this type of testing is preferred on condition that MIMO model estimation is also possible. This will be further discussed in Section 4.5.4.

The ASYM method, developed by Zhu [4], ensures informative data by making use of optimal test signals designed to minimise the sum of the squares of the simulated error. The desired spectra of these test signals are realised by PRBS signals or filtered white noise.

An approach to simultaneous constrained MPC and Identification (MPCI) was developed by Shouche *et al.* [2]. In this approach, a persistent identification criterion is used as an additional constraint in the standard on-line optimisation of MPC.

Case studies were done by applying (1) MPCI, (2) MPC with external dithering signals (PRBS signals were added to control signals) and (3) MPC with no external input (constant references). The first two cases gave good results due to the process excitation by the controller in the MPCI case, and the external dithering signals in the MPC case. However, the case study of the MPC with no excitation, gave very poor parameter estimates due to the lack of information about the dynamics of the process [2].

Therefore, although the MPC controller is nonlinear, time-varying and complex, which in general yield informative experiments [11], a persistently exciting signal $r(t)$ will guarantee this. The type of excitation signal needed, depends on the characteristics of the plant. For choice of excitation signals, refer to Section 3.8.

In the case of the MIBK plant, data sets from the normal operation of the system were used for the identification of the plant. Since no knowledge of the reference signals was available, it is possible that these signals were not PE and that the system is, therefore, not identifiable from this data.

However, another option is to inter-sample the plant inputs and outputs, since this is also claimed to ensure identifiability - even when there is no persistent excitation signal [18]. The question of identifiability has only a *yes* or *no* answer. Another important question is how precise the model identified from the inter-sampled data is. The model may be identifiable, but still have such a large variance that it is unacceptable [36], as illustrated in the variance simulation study in the next section.

4.4 Evaluation of the Inter-Sampling Approach

A simulation study was done to determine how precise a model identified from closed-loop inter-sampled data is. In this study the variance of the identified model was evaluated. An estimation of the variance was obtained from a Monte Carlo simulation.

4.4.1 Set-Up of the Variance Simulation

In the Monte Carlo simulation the closed-loop system, described in Addendum A, was simulated to determine the closed-loop response signals, with the seed of the added noise different in each run. A total of 50 runs were performed. For each run in the simulation the resulting input and output signals of the plant were logged. From each of these fifty sets of input-output signals a plant Δ-model, described in Section 3.9.2, was identified with the direct closed-loop SID approach and the PEM estimation method. From these Δ-models the T-models, also described in Section 3.9.2, were determined, making use of the *d2d* MATLAB function that implements Eqn. (3.74).

From these models the estimation error was determined with Eqn. (4.1)

$$ERR = \frac{1}{50}\sum_{l=1}^{50}\left(\left|\hat{a}_i^{(l)} - \bar{a}_i\right|^2 + \left|\hat{b}_i^{(l)} - \bar{b}_i\right|^2\right), \tag{4.1}$$

where $\hat{a}_i^{(l)}$ and $\hat{b}_i^{(l)}$ are the parameter estimates for run l of the simulation and $\bar{a}_i$ and $\bar{b}_i$ are the mean parameter estimates.

The variance of the identified models was also evaluated, by determining the variance in the step, impulse, Bode magnitude and Bode phase responses of the fifty identified models. The variance of the step responses (*step_var*) was determined making use of Eqn. (4.2) and the *var* function in MATLAB, which determines the variance, i.e. square of the standard deviation for a set of values.

$$step\_var = \frac{1}{L}\sum_{t=1}^{L} var(step\_resp(t)), \text{ where} \tag{4.2}$$

$$step\_resp(t) = \begin{bmatrix} step^{(1)}(t) \\ step^{(2)}(t) \\ \vdots \\ step^{(50)}(t) \end{bmatrix}, \text{ and where}$$

$step^{(l)}(t)$ is the step response value at time index t of the model, identified in run l of the simulation, and L is the time span over which the step responses was determined. Eqn. (4.2) determines the variance of the step response at each time instant t and then determines the average variance from these values. The variance of the impulse responses was determined similarly. Also, the variances of the Bode magnitude and phase responses were determined similarly, with t substituted with the frequency index and L substituted with the frequency span of these responses.

A number of cases are evaluated in which different parameters were varied:

- case 0 (verification case): the reference signal was stepped to 10 and the integer p, which determines the number of times the signals are inter-sampled, was varied,

Reference signal equal to zero:

- case 1: p was varied,
- case 2: similar to case 1, p was varied , but the number of samples N was kept constant,
- case 3: p was kept constant, but the noise power was varied,

Reference signal stepped from 0 to 0.001:

- case 4: p was varied,
- case 5: p was kept constant, but the noise power was varied,
- case 6: p was kept constant, but the bandwidth of the plant was varied (a different closed-loop system, also described in Addendum A was used), and
- case 7: for the plant, simulated in open-loop, p was varied.

For these case the variances and estimation errors were determined by doing the Monte Carlo simulation for different values in these parameters.

In case 0 (verification case), identifiability was ensured by stepping the reference signal to a significantly larger value. This case verifies the software: it shows that under normal conditions, i.e. PE reference signal and a good SNR, the software can be used to identify an accurate and precise model.

In the first number of cases, the reference signal was made zero to ensure that this signal is not PE. In these cases, p was varied to determine how the precision of a model is influenced

Table 4.1: Variance in Responses and Estimation Error in Parameters

	Step	Impulse	Bode Magnitude	Bode Phase	ERR
$p = 1$	4.9108×10^{114}	2.0656×10^{115}	52.3338	3.2036×10^{3}	0.0835

by the number of times the output is inter-sampled, when identifiability is not ensured with a PE reference signal or a nonlinear controller. The number of samples N was then kept constant to see if the influence of p is only due to the increase in data samples. Furthermore, the noise power was varied to determine what the influence of the SNR is on the precision of the model when the reference signal is zero.

In the other cases the signals were stepped from zero to a very small value to ensured a PE reference signal and a large (bad) SNR. In these cases p was varied to determine how the precision of a model is influenced by the number of times the output is inter-sampled, when identifiability is ensured with a PE reference signal, but with the SNR very large. The noise power was also varied to determine what the influence of the SNR is on the precision of the model. Lastly, the bandwidth of the plant was varied to determine if the influence of p is dependent on the sampling frequency $\frac{1}{T}$ relative to the plant bandwidth. The influence of p was also evaluated in open-loop in order to verify the closed-loop results.

4.4.2 Results of Variance Simulation

4.4.2.1 Case 0 (Verification Case)

In the verification case a satisfactory, i.e. accurate and precise, model was identified for any value of p. The identified model agrees very well with the true model in both the time domain, i.e. step and impulse responses, and frequency domain, i.e. phase and magnitude responses. These results verify the simulation software for the variance analysis.

4.4.2.2 Case 1

When $p = 1$ the data were not inter-sampled, i.e. the plant output was sampled at the same rate as the control input. For $r(t) = 0$ and $p = 1$ unstable models were identified and the estimation error, as well as the variance in the step, impulse, Bode magnitude and Bode phase responses are large, as can be seen in Table 4.1.

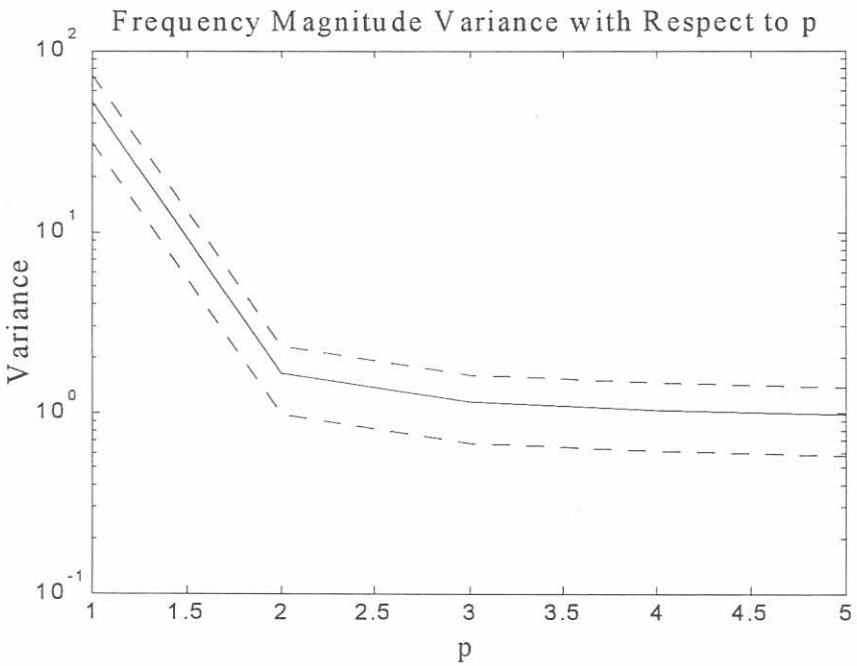

Figure 4.1: The variance in the Bode magnitude response for different values of p. The dashed lines represent the 95% confidence bounds.

For $p = 2$ most of the identified models were stable and the estimation error, as well as the variance in the step , impulse and Bode magnitude, are significantly smaller, as shown for the magnitude response in Fig. 4.1. Figure 4.2 shows that the reduction in variance for the Bode phase response only starts to happen at $p = 3$. The variance as well as the bias in the model are still very large compared to values obtained in the verification case.

For $p > 2$, the estimation error and variances for most of the responses are slightly reduced from those obtained for $p = 2$. Figures 4.1 and 4.2 show how the variances are reduced for the Bode magnitude and Bode phase responses. In the figures the 95% confidence bounds of the determined *variance values* are also shown. These confidence bounds are determined as in [42]:

$$\text{confidence bound} = \textit{variance value} \pm 2 * \sigma, \quad \text{where} \tag{4.3}$$

$$\sigma = \sqrt{\frac{2}{\text{simulation runs} - 1}} * \textit{variance value}, \quad \text{simulation runs} = 50. \tag{4.4}$$

4.4.2.3 Case 2

The results obtained in case 2 are similar to the results of case 1. However, the estimation

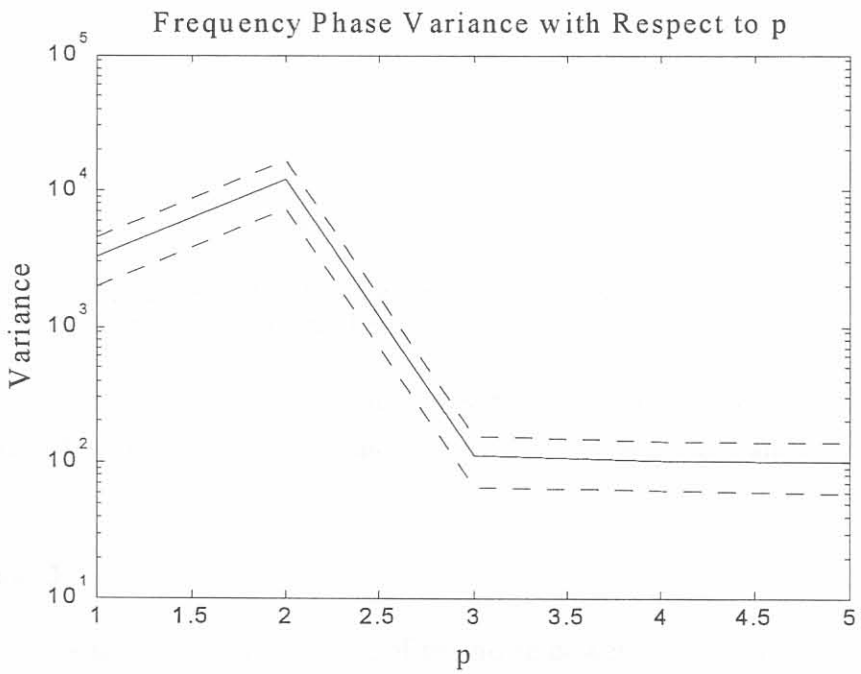

Figure 4.2: The variance in the Bode phase response for different values of p. The dashed lines represent the 95% confidence bounds.

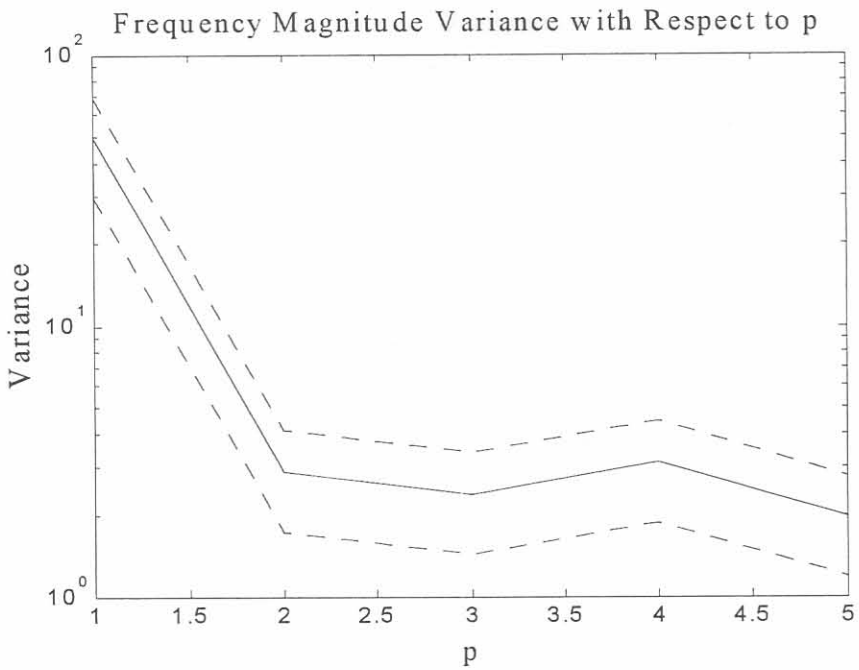

Figure 4.3: The variance in the Bode magnitude response for different values of p, with N constant. The dashed lines represent the 95% confidence bounds.

error and variances do not decrease as much as in case 1, but stay almost constant at higher values. Figure 4.3 shows how the variance of the Bode magnitude change.

4.4.2.4 Case 3

For a constant p and $r(t) = 0$, the size of the noise power does not have any effect on the size of the variances for the step, impulse, Bode magnitude and Bode phase responses.

4.4.2.5 Case 4

In case 4 the estimation error and variances of the responses are much smaller than in case 1, but still unsatisfactory. There is also not a big difference between the variance for $p = 1$ and $p \geq 2$. Again, the variance in the Bode magnitude response is shown in Fig. 4.4. For most of the responses the variances stay almost constant or even increase.

4.4.2.6 Case 5

When the reference was stepped, there was a small transient effect in the plant output. For

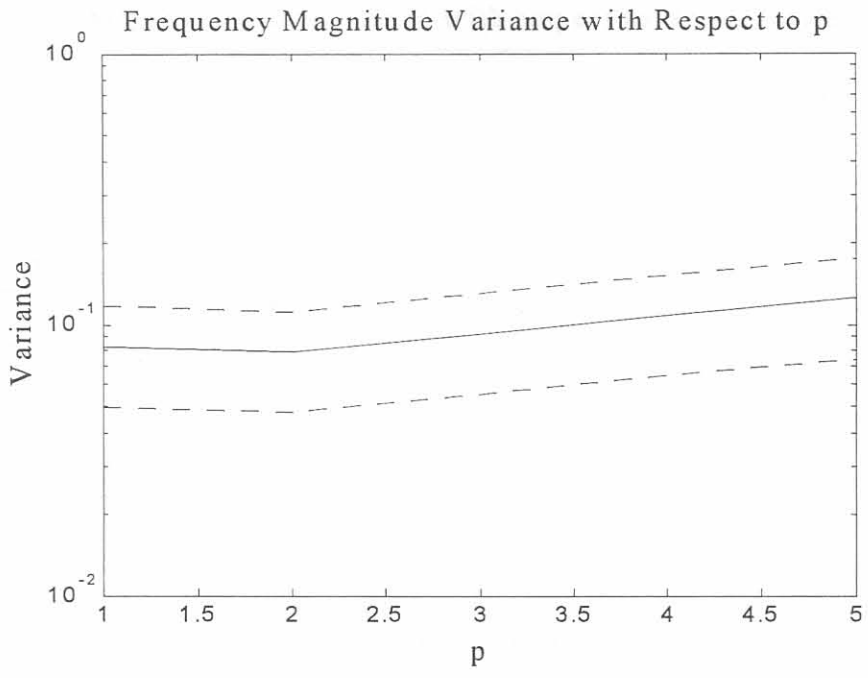

Figure 4.4: The variance in the Bode magnitude response for different values of p, with the reference signal stepped to 0.001. The dashed lines represent the 95% confidence bounds.

a very small noise power the transient effect was still visible, which resulted in a small variance, i.e. a precise estimation of the plant. As the noise power increases, the SNR worsens and the variance increases, as shown in Figs. 4.5 and 4.6. In Fig. 4.5 the 95% confidence bounds are too large to be shown.

4.4.2.7 Case 6

The original sampling frequency was made much slower than ten times the plant bandwidth. Repeatedly, the plant bandwidth was decreased by changing the time constant, and a new controller was designed (see Addendum A), but the execution time of the controller, as well as original sampling time T, was kept constant. The estimation error and variances of the responses for different values of p were determined.

When the original sampling frequency is much smaller than ten times the plant bandwidth, an increase in p decreased the estimation error, as well as the variances in the responses. However, as the original sampling frequency comes close to ten times the plant bandwidth, an increase in p does not significantly affect the variance any more.

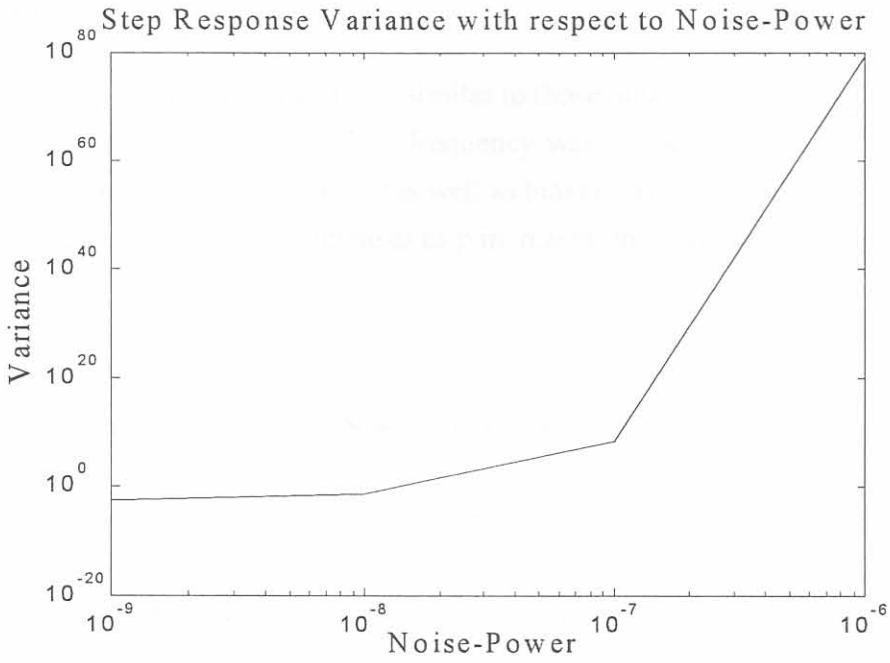

Figure 4.5: The variance in the step responses, with respect to the noise power, for a small step in the reference input.

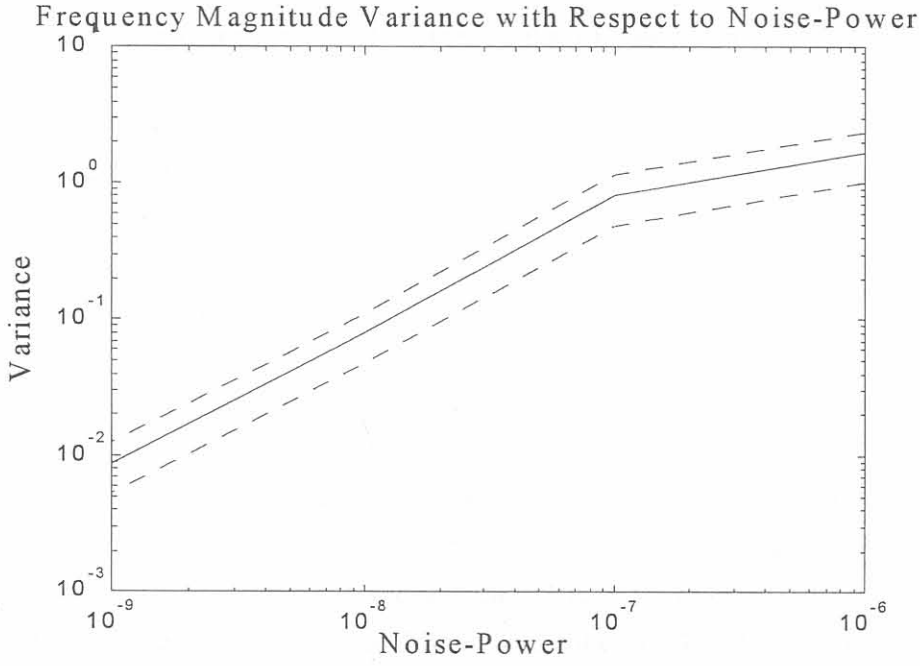

Figure 4.6: The variance in the Bode magnitude responses, with respect to the noise power, for a small step in the reference input. The dashed lines represent the 95% confidence bounds.

4.4.2.8 Case 7

For the open-loop system, the results are similar to those obtained in cases 4, 5 and 6. Firstly, N was kept constant and the sampling frequency was in the range of ten times the plant bandwidth. Again, the model variance (as well as bias) is large and an increase in p does not decrease the variance. When N increases as p increases, the variance improves only slightly [36].

4.4.3 Discussion of the Variance Simulation Results

From the fact that the estimation error and variances in case 1 decrease significantly from $p = 1$ to $p = 2$ or 3, it is concluded that even when the controller is of a lower order than the plant and the reference is not excited, inter-sampling the input and output of the plant at least once ($p = 2$) does make the plant identifiable. However, these results also show that identifiability does not guarantee a small variance or a small bias. Furthermore, this case shows that an increase in p does not have a significant effect in decreasing the variance. The small decrease in the variance can mostly be contributed to the fact that the number of samples N increases as p increases, because in case 2 where N is kept constant, the variance does not decrease as much as in case 1, but stays almost constant at a higher value.

For some responses an increase in variance occurs at certain instances. These increases can be contributed to the convergence error from the Δ-model to the T-model that increases with p [18]. For p very large, i.e. a small Δ, an increase in the variance can also be expected, because of the fact that a very small sampling time causes most least squares algorithms to become numerically unstable [11].

For a zero reference signal, there is no transient effect present in the plant output. Case 3 shows that in this case an increase or decrease in the noise power does not effect the variance and the variance is, in general, very large. The SNR refers to the ratio between the plant input signal and the noise. The noise power has no effect, since in this situation, the input signal only originates from the noise and thus the SNR is not relevant. However, when even a very small test signal is added tot the reference, as in cases 4 and 5 where the reference is stepped from 0 to 0.001, the variance is much smaller than obtained in cases 1, 2 and 3. In this situation the size of the noise power and thus the SNR has an effect on the size of the variance and the variance is only acceptable when the SNR is good. This agrees with the variance expression in Eqn. (3.70), where it is the signal-to-noise ratio, where *signal* is the

part of the plant input that derives from the injected reference, that determines the quality of the open-loop transfer function estimate.

Case 6 shows that when the original sampling frequency $\frac{1}{T}$ is very slow in comparison with the plant bandwidth, i.e. much less than ten times the plant bandwidth (e.g. 50 times), then an increase in p will decrease the variance significantly. This agrees with the rule that the sampling frequency should lie in the range of ten times the plant bandwidth [11].

The open-loop simulation in case 7 confirms the closed-loop results, namely that an increase in p does not decrease the variance significantly.

4.4.4 Conclusion of the Variance Simulation

It is concluded that, although the inter-sampling approach ensures identifiability, it does not ensure a precise model with a small variance (or bias). Structured tests, where external test signals are added to the reference inputs, should be performed to ensure good SNRs.

Also, increasing the sampling frequency more than twice from the control action rate, will only reduce this variance significantly if the original sampling frequency $\frac{1}{T}$ is much slower than ten times the plant bandwidth.

4.5 Model Structure Selection

The choice of a suitable type of model structure is a crucial step in the identification process in order to obtain a good and useful model [6]. For closed-loop identification, the choice of the model structure depends on three often conflicting issues [4]:

- the consistency of the model in closed-loop identification,
- the compactness of the model, and
- the numerical complexity in estimating the parameters.

These and related issues will now be discussed.

4.5.1 Model Consistency

Since the direct SID approach is chosen in Section 4.2, the only estimation method that delivers consistent results when implemented on closed-loop data, is the PEM estimation method. With the PEM approach, the models may have arbitrary parameterisation.

The following model structures are all special cases of the more general PEM model family:

- FIR,
- ARX,
- ARMAX,
- OE, and
- BJ.

For all of these structures, the parameters are determined by minimizing the sum of squares of the prediction error [3]. Therefore, any of these structure will be consistent in direct closed-loop SID, except OE that will fail when the input noise is not white [11].

As already explained in Section 3.4, OE models, as well as the nonparametric models obtained from instrumental variables, spectral analysis and many subspace methods, are not consistent in closed-loop.

4.5.2 Compactness and Parametric Structures

Parametric models, such as the ARX and the ARMAX, are much more compact, since they need fewer parameters to describe a plant's dynamic behaviour than nonparametric models, such as FIR models (according to Zhu *et al.* [41] FIR models are nonparametric, but some may argue that FIR models are parametric). A more compact model will be more accurate and precise, provided that the parameter estimation algorithm converges to a global minimum and the model order is selected properly. Zhu *et al.* [41] show that the bias of parametric models is smaller than the bias caused by truncation in FIR models, since parametric models have infinite length (e.g. infinite impulse response) and no truncation is necessary. Parametric models also reduce the model variance of nonparametric models.

Also, for the same model precision, a parametric model requires up to 75% less testing time when compared to nonparametric models; or, put in another way, for the same test data, a parametric model can be much more precise [41].

Therefore, a parametric model is a better choice in the modeling of industrial processes, which are the kind of processes that are controlled by MPC controllers. Since many of these industrial processes may also be unstable, it seems natural to choose from these parametric

models the ARX or ARMAX model structure, because these models give stable predictors [11].

4.5.3 Numerical Complexity

As stated in Section 3.7.4, when the ARX and FIR models are used together with a quadratic criterion function, the standard linear LSE method is obtained. This means that parameter estimates can be found by solving only a standard LSE problem, which can be done analytically. Typically, for other parameterisations and criteria, one has to rely on some iterative search scheme to find the best estimate [6].

The linear LSE method is numerically simple and reliable, and problems with local minima or convergence do not occur [41].

Although the ARMAX model is more compact than the ARX model, the numerical complexity in the parameter estimation of the ARMAX model is much higher than for the ARX model, since nonlinear optimisation routines are needed. Therefore, the ARMAX optimisation routines often suffer from local minima and convergence problems when identifying multivariable processes [4].

4.5.4 Multivariable Models

Multivariable systems are often more challenging to model [11]. As mentioned, data from a single variable test may not contain sufficient information about the multivariable character of the process and multivariable testing can solve this problem [3]. However, in multivariable testing all the reference signals can be excited simultaneously. This makes SISO model estimation, where only the parameters of one SISO transfer functions can be estimated at a time, undesirable. Multiple-Input-Single-Output (MISO) model estimation, where partial models of a system's behaviour are constructed from all the input channels and one output channel, is more desirable, but MIMO model estimation of all the transfer functions is the most desirable. The reason for this is that models for prediction and control will be able to produce better results if constructed for all outputs simultaneously. This follows from the fact that knowing the set of al previous output channels gives a better basis for prediction than just knowing the past outputs in one channel [11].

Multivariable ARX (IDARX) models in MATLAB are parametric models with several inputs and several outputs that can be estimated using a standard MATLAB function. Here

it is allowed for n_a, the order of polynomial $A(q)$, n_b, the order of polynomial $B(q)$ and n_k, the time delay, to contain one row for each output number and one column for each input number. For most other model structures several MISO models have to be identified, which can then, afterwards, be combined into a multivariable model [43].

4.5.5 Model Order Selection

For the purpose of control, it is important to select the model order so that the process model from inputs to outputs is accurate. In the time domain, this requires for the simulation error or predicted output error of the model to be minimal [4]. See Section 4.6.1.1 for a discussion of the simulation error and predicted output error.

For models of the ARX type, various orders and delays can be efficiently studied with the command *arxstruc* in MATLAB. The ARX model can be fitted to the validation data set for many different model structure orders. For each of these models, the sum of squared prediction errors is computed, as they are applied to the validation data set. The structure (order) that has the smallest loss function (best fit) for the validation set can then be selected. Such a procedure is known as *cross-validation* and is a good way to approach the model order selection problem. It is usually a good idea to visually inspect how the fit changes with the number of estimated parameters [43].

A good idea is to first establish a suitable value for the delay by testing second-order models with different delays. Use the delay that yields the best fit. All combinations of ARX models, with different orders for polynomials $A(q)$ and $B(q)$, with delays around the chosen value can be inspected to make sure [43].

However, if the model is validated on the same data set from which it is estimated the fit always improves as the flexibility of the model structure increases. One needs to compensate for this automatic decrease of the loss functions. There are several approaches. The best known technique is Akaike's Final Prediction Error (FPE) criterion and his closely related Information Theoretic Criterion (AIC). Both simulate the cross-validation situation, where the model is tested on another data set [11].

The FPE is formed as

$$FPE = \frac{1+\frac{d}{N}}{1-\frac{d}{N}}V, \tag{4.5}$$

where d is the total number of estimated parameters and N is the length of the data record. V is the loss function (quadratic fit) for the structure in question. The AIC is formed as

$$AIC = \log(V(1+2\frac{d}{N})). \tag{4.6}$$

According to Akaike's theory [43], the model with the smallest FPE, or AIC should be chosen.

A related criterion is Rissanen's Minimum Description Length (MDL) approach, which selects the structure that allows the shortest over-all description of the observed data [43].

If substantial noise is present, the ARX models may also need to be of high order to describe simultaneously the noise characteristics and the system dynamics. The reason being that for ARX models the disturbance model $1/A(q)$ is directly coupled to the plant model $B(q)/A(q)$ [43].

There are different ways to go about selecting the model order. According to Forssell [6], one typically starts by checking if the model of lowest order is sufficiently good. The model order is then increased until a model that passes the validation test is found. If the resulting model order is too high, since a high order is often not desirable for control design [44], model reduction can be considered. As a general rule of thumb, one then knows that the variance error will dominate the bias error for this model [6].

However, Leskens *et al.* [44] first identify a very high-order ARX model and then, afterwards, reduce the model order. Zhu *et al.* [3] also use this method in the ASYM approach.

It is decided to follow the more typical approach of first identifying a lower order model and then increasing the order, depending on the validation results.

4.6 Model Validation

Before a model can be delivered to the user, it has to pass some validation test. Model validation can loosely be said to deal with the question of whether the best model is also *good enough* for its intended use [6].

What is a good model? In the linear case, the answer is that the model fit has to be good in certain frequency ranges, typically around the cross-over frequency. In general, one can

say that the desired control performance dictates the required quality of the model fit: the higher the demands on control performance are, the better model fit is required [6].

The methods chosen for the validation of the identified model are discussed under different headings. Some of the validation methods form part of the validation step in the proposed methodology and other methods are used to validate the methodology itself:

- the methods that fall under *standard validation* form part of the proposed methodology and can always be used to validate the identified models, and
- the methods that fall under *comparison with open-loop identified model* and *examination of closed-loop system* are used to validate the proposed methodology and does not form part of the methodology itself.

4.6.1 Standard Validation with the Closed-Loop Data

Common validation tools are *residual analysis* and *cross-validation*, where the model is simulated using *validation* data and where the output is compared to measured output data [6]. The *validation* data are a set of the measured closed-loop data that is different from the *estimation* data set. The aim is to validate the model using data that were not used to fit the model, since it is not so surprising that such a model will reproduce the estimation data. Therefore, this is a much more stringent test than using the same data for fitting and validation [11].

If a limited amount of data is available about two thirds of the region should be used to fit the data and one third should be used to validate the model.

4.6.1.1 Simulation and Prediction

In the simulation and prediction test, it is determined how well the model is capable of reproducing the validation data. One can work with k-step ahead model predictions $\hat{y}_k(t \mid m)$ as the basis for comparison. This means that $\hat{y}_k(t \mid m)$ is computed from past data,

$$u(t-1), \ldots u(1), y(t-k), \ldots, y(1), \tag{4.7}$$

using the model m. The case when k equals ∞ correspond to the use of past inputs only, i.e., pure simulation [11]. This is a very stringent test.

For control applications, the predicted output over a time span corresponding to the dominant plant model time constant, will be an adequate variable to look at [11]. In the case of MPC controllers, the model output is predicted over a model horizon [30]. Thus, the model has to predict the output accurately, i.e. reproduce the validation data, only for a number of steps ahead, equal to the model horizon h. A model that satisfies this condition will be satisfactory for the design of an MPC controller. In re-identification the model horizon of the old controller is known and this value can be used for validation.

The model can then be evaluated by a visual inspection of plots of $y(t)$ and $\hat{y}_k(t \mid m)$, or by the numerical value of fit $J_k(m)$ for N number of samples:

$$J_k(m) = \frac{1}{N} \sum_{t=1}^{N} |y(t) - \hat{y}_k(t \mid m)|^2 . \tag{4.8}$$

4.6.1.2 Residual Analysis Test

The *leftovers* from the modeling process - the part of the data that the model could not reproduce - are the *residuals*,

$$\varepsilon(t) \quad = \quad y(t) - \hat{y}_k(t \mid m). \tag{4.9}$$

These contain information about the quality of the model and can by analysed to draw conclusions about the validity of the model [11]:

Whiteness Test: An *auto-correlation* of the error signal, $R_\varepsilon^N(\tau)$, determines whether the error signal is white noise or not. If it is white then the model is an unbiased estimator, which means that model parameter estimates will be *true* on average [11]. If the auto-correlation function of the error signal does not fall significantly outside the 99% confidence region, except for $R_\varepsilon^N(\tau) = 1$, then one can reasonably assume that the error is white noise. The errors can also be plotted and visually inspected to see if outliers are present or if it represents white noise.

Independence of Residuals and the Past Inputs: The independence of residuals and past inputs can be determined from the cross-correlation between these signals, $R_{\varepsilon u}^N(\tau)$. An appealing way to carry out this test is to plot the *cross-correlation* function. The confidence limits of this function will be horizontal lines. For a good model, the cross-correlation function should not fall significantly outside the 99% confidence region. If negative correlation

is present, it does not mean that the model structure is deficient, only that output feedback occurs - the current error influences the future input [11]. A rule of thumb is that a slowly varying cross-correlation function outside the confidence region is an indication of too few poles, while sharper peaks indicate too few zeros or incorrectly estimated delays [43].

If these cases are satisfied, the model is, statistically speaking, a very good model.

4.6.1.3 Model Reduction

One method that tests if a model is a simple and appropriate system description, is to apply some model reduction technique to it. If a model order can be reduced without affecting the input-output properties very much, then the original model is unnecessarily complex [11]. Söderström [45] has developed this idea for pole-zero cancellation. If the pole-zero plot, including confidence intervals, indicate pole-zero cancellation in the dynamics, this suggests that a lower order model can be used. In the case of ARX when pole-zero cancellations occur, the extra poles are usually introduced to describe the noise. In this case another model structure should be tried [11].

4.6.2 Comparison with the Open-Loop Identified Model

The proposed closed-loop SID methodology will be used for re-identification of the plant. Since the open-loop SID method was previously used, the closed-loop identified models should be compared to models obtained from open-loop tests.

To validate the proposed identification methodology, the plant model in question should, therefore, also be estimated from data obtained from open-loop tests. If the model identified from the open-loop test data is accurate and precise, it can serve as a reference model for evaluating a model obtained from the closed-loop data. If the models are comparable, it is an indication that the proposed identification methodology is valid.

Since the models are black-box models, the consistency of the input-output behaviour of the models should be evaluated [11]. This can be done as follows:

Firstly, these models can be compared visually in the time and frequency domain. The Bode phase and magnitude responses as well as the step and impulse responses of both models can be compared. Similarly the pole-and-zero plots and Nyquist plots can be compared.

Secondly, the fit of the measured output with the pure simulated output and with the h-step predicted output of the open-loop identified model can be compared with the closed-loop identified model's fit.

Lastly, a residual analysis test can be done for the open-loop identified model. These results can then be compared with the results obtained for the closed-loop identified model.

These tests should be performed using the same validation data set. A good comparison will indicate that the identification methodology is comparable in accuracy to open-loop SID.

It is possible that the closed-loop identified model can be more accurate than the open-loop identified model. Therefore, in the simulation case where the true model is known, the open-loop identified model and the closed-loop identified model should also be compared, by evaluating how close these models are to the true model. The accuracy of these models can be determined by visually and numerically comparing them to the true model in the time and frequency domain. A typical numerical value that can then be compared for the two estimated models is the relative norm

$$
\begin{aligned}
freqfit &= \left\| \frac{\left\| \hat{G}_N(\omega) \right\|_2^2 - \|G_0(\omega)\|_2^2}{\|G_0(\omega)\|_2^2} \right\|^2, \quad \text{where} \qquad (4.10)\\
\|G(\omega)\|_2^2 &= \sup_x \frac{\|G(\omega)x\|^2}{\|x\|^2}, \quad \text{and}\\
\|x\|^2 &= \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}.
\end{aligned}
$$

The model that gives the smallest value for Eqn. (**??**) has the best fit in the chosen frequency range. A similar value can be computed in the time domain, e.g. the relative norm for the step response coefficients

$$
stepfit = \left(\sum_{i=1}^{n} \left| \frac{\hat{s}_i - s_{0i}}{s_{0i}} \right|^2 \right)^{1/2}. \qquad (4.11)
$$

4.6.3 Examination of the Closed-Loop System

The goal of model validation is to test whether the model is good enough for its purpose and to provide advice for possible re-identification if the identified model is not valid for

its intended use [3]. The purpose of the model in question is to design an acceptable MPC controller. Therefore, the final proof rests in demonstrating the acceptable performance of the controller.

A good validation test is therefore to see how the MPC controller, designed from the closed-loop identified model, controls the true plant. If the regulator, based on the identified model, gives satisfactory control, the model is a *valid* one [11].

Often, it will be impossible, costly, or dangerous to test all models with respect to their intended use in practice [11]. Therefore, this validation method is not included in the proposed methodology. However, when the true model is known, a good validation test of the methodology is to simulate how the MPC controller, designed from the identified model, controls the true plant. The stability of the resulting closed-loop system should be analysed. The system is stable if all the poles of the discrete closed-loop system are inside or on the unit-circle.

4.7 System Identification Toolbox

Nowadays there are well-supported and user friendly tools available for the identification of linear systems on the basis of experimental data. In particular, the Mathwork's System Identification Toolbox (SITB), version 6.1, which is equipped with a graphical user interface, can be mentioned. This toolbox enables the user to identify and validate models in different types of model structures. Additionally, there is users' support in terms of graphical tools for model evaluation, as well as support for e.g. bookkeeping of identified models [20].

In this SITB there are only limited possibilities to identify models on the basis of data that are obtained under closed-loop experimental conditions. This particular experimental situation often requires special treatment, in the sense that besides input and output signals of a plant, measured external excitation signals can be involved, as well as some, possibly known, controllers that are implemented on the system. In these cases an add-on to the SITB, such as the CLOSID toolbox is needed [20].

Luckily, in the direct closed-loop SID approach no reference signals, or controller information, are involved in the identification process, only the plant inputs and outputs. Thus, the SITB, without any custom-written algorithms or extra software, is sufficient and suitable for the direct closed-loop SID approach. Therefore, no new software has been developed for this methodology.

The functions of the MATLAB SITB, used in the implementation of the methodology, are discussed in Section 5.3 and Section 6.2.

4.8 System Identification Steps

The appropriate closed-loop SID methodology, for plants controlled by MPC controllers, can now be summarised by addressing each of the five SID subproblems, discussed in Section 3.2. Each of these subproblems can be seen as a step in the SID procedure. The five steps, together with a summary of the chosen methods for validation of the methodology, are given.

The direct closed-loop SID approach should be used when implementing these SID steps. This approach will deliver the best SID results for plants controlled by nonlinear controllers, such as constrained MPC controllers, since this method is applicable to systems with arbitrary feedback mechanisms.

4.8.1 Experiment Design

Experiment design involves issues like choosing what signals to measure, choosing the sampling time, and choosing excitation signals. Once these issues have been settled, the actual identification experiment can be performed and process data be collected.

4.8.1.1 Signals to be Measured

Since the direct closed-loop SID approach is used, only the inputs and outputs of the plant have to be measured. Since this methodology will usually be implemented for re-identification purposes, the signals to be used as inputs and outputs would have been chosen earlier on and would therefore be known at this stage.

4.8.1.2 Sampling Time

The sampling frequency should lie in the range of ten times the plant bandwidth [11]. Since the MPC controller is already implemented, the measurement devices are usually already in place and the sampling time predetermined. Usually, the sampling time of these devices is the same as the execution time of the controller.

If possible, the inputs and outputs should be inter-sampled in such a way that it is sampled twice as fast as the execution time of the controller. Although the identified model may still have a very large variance, at least identifiability will be ensured when applying the inter-sampling method.

4.8.1.3 Excitation Signals

If possible, one should always try to add a persistently exciting signal to the reference input, either $r_a(t)$ or $r_b(t)$, since this is the easiest way to ensure identifiability. The type of signal is determined by the plant dynamics.

4.8.2 Data Collection

4.8.2.1 Collection

The collection of the data should be straightforward, since in the case of industrial plants, controlled by MPC controllers, the inputs and outputs of the plant are constantly logged. Usually there is some type of database present from which the desired data can be retrieved.

4.8.2.2 Preprocessing

The raw data that have been collected from identification experiments are not likely to be suitable for immediate use in identification algorithms. There are several possible deficiencies in the data that should be attended to:

- high frequency (above the frequency of interest to the system dynamics) disturbances in the data record,
- occasional burst and outliers, missing data and non-continuous data records, and
- drifts and offsets, low-frequency disturbances, possibly of a periodic character [11].

In off-line applications, the data sets should always be plotted first to inspect them for these deficiencies. Standard preprocessing methods used for open-loop data can also be used for closed-loop data. In the MATLAB SITB there are many routines to plot data, filter data, and remove trends in data. The preprocessing methods used in this specific application are discussed in Section 5.3 and Section 6.2.

Since a prediction error model is employed, it is especially important to remove trends, drifts and outliers. This will prevent the discrepancy in signal levels to dominate the criterion of fit, which could mask the dynamic properties [11].

4.8.2.3 Time Delay

At this stage, the time delay of the model should also be selected. This can be estimated by making use of visual inspection of the data inputs and outputs as well as knowledge of the plant and the previous model. After the model has been estimated, the choice of time delay can be validated. If the validation results, especially the residual analysis, is unsatisfactory, the time delay should be re-estimated and the model re-identified. It is easy to estimate many models for different time delays in the MATLAB SITB. From these models the one with the best fit can be selected.

4.8.3 Model Structure Selection

4.8.3.1 Type of Structure

The ARX type model structure is the best choice of model structure, provided that the noise model is accurate for the process to be modelled, which will usually be the case in re-identification.

This step should be done with care. The plant in question should always be considered carefully before a final choice of the model structure is made. If the ARX structure does not deliver good validation results, then other structures can easily be determined and compared, making use of the standard validation methods in MATLAB.

4.8.3.2 Order Selection

The selection of the model order forms part of the model structure selection. In the case of re-identification an older version of the model is usually available. The order of the older model can be used - at least for a start. After the model has been estimated, this choice of order can be validated. If the validation results are unsatisfactory another order should be chosen and the model re-identified. Actually, many models for different orders, from low to

high, can then be estimated and the best order can be selected by comparing loss functions. In the case where no validation data are available, one should make use of the AIC, FPE or MDL model structure selection criteria. The procedure is described in Section 4.5.5.

4.8.4 Model Estimation

Given a suitable model structure and measured data, one can turn to the actual estimation of the model parameters. Since there exist special-purpose software tools that are very efficient and easy to use for model estimation, this step is perhaps the most staightforward one [6].

The estimation method should, of course, be the PEM estimation method, since it delivers consistent results for closed-loop data. A multiple-output model can be estimated with the standard MATLAB SITB command, *idarx,* and a MISO or SISO model can be estimated with the *arx* command. These function use the standard LSE method to estimate the best parameters for an ARX model structure.

4.8.5 Model Validation

For model validation the standard validation methods, outlined in Section 4.6, together with a measured closed-loop validation data set, not used in the model estimation step, can be used. These are:

- simulation and prediction,
- residual analysis, and
- model reduction - check for pole-zero cancellation.

If the model fails the validation test, some, or all, of the above steps have to be reconsidered and repeated until a model that passes the validation test is found.

4.8.6 Methodology Validation

The methodology validation step does not form part of the proposed methodology itself, but is included, because, in this work, it forms the final step in the selection of an appropriate methodology. In this step the methodology is validated. In the case where the results obtained from this step are unsatisfactory, the whole methodology should be re-evaluated. The methods employed for the validation, also outlined in Section 4.6, are:

- a visual comparison of the closed-loop identified model with the open-loop identified reference model in both the time and frequency domain,
- comparison of the pure simulation and prediction fit of the closed-loop identified model with the fit obtained for a model identified in open-loop,
- comparison of the residuals of the closed-loop identified model with those obtained for a model identified in open-loop,
- a comparison of the open-loop and closed-loop identified models, by evaluating how close these models are to the true model, and
- examination of the closed-loop system's stability for the controller designed from the new model.

4.9 Conclusion

By reviewing the relevant literature one can see that there are many options available, regarding identification approaches, guarantees of identifiability, model structures and model validation techniques. From all these options the most appropriate ones were chosen to be used in the closed-loop identification of plants controlled by MPC controllers. These choices are mainly based on: the theory regarding closed-loop SID; characteristics of MPC controllers, e.g. nonlinearity; characteristics of industrial plants, e.g. multivariable; keeping the methodology relatively uncomplicated; and results obtained by other researchers for similar cases.

The methodology was, thus, developed by making selections, based on certain criteria, from the many available options in closed-loop system identification.

Closed-Loop SID Approach: The direct closed-loop SID approach was chosen for the proposed methodology. The direct approach is actually the only option for the MIBK process, since the reference signals, as well as the controller settings, of the MIBK process are not available and the indirect and joint input-output approaches can, therefore, not be implemented. However, it is concluded that the direct approach will deliver the best SID results for plants controlled by nonlinear controllers, such as constrained MPC controllers, since this method works regardless of the complexity of the controller. This approach also ensures consistency and optimal precision and simplifies the development of the methodology, since it makes use of the standard functions in the MATLAB SITB and does not require any custom-written algorithms or extra software.

Guarantee of Identifiability: It can be concluded that, although the MPC controller is nonlinear, time-varying and complex, which in general yield informative experiments, this is not a guarantee for identifiability. The reason being that the changes in a multivariable MPC structure, as it deals with changes in the active constraints, are unpredictable under normal operation and can thus not guarantee a given number of changes in the controller settings. A PE reference signal will, however, guarantee identifiability.

Inter-Sampling: From the variance simulation study on the inter-sampling approach it is concluded that, although this method ensures identifiability, without structured tests that ensure good SNRs, the variances of the identified models are very large and the precision is thus unsatisfactory.

Model Structure: The ARX type model structure was chosen. This is the best choice of model structure, provided that the noise model is accurate for the process to be modelled, since:

- it utilises the consistent closed-loop PEM estimation method,
- it is parametric, which in turn ensures compactness and accuracy,
- it can handle unstable systems without problems, since the predictor is stable,
- its parameters can be determined from a numerically simple and reliable LSE method that does not suffer from problems with local minima or convergence, and
- MATLAB allows for ARX structures to be estimated from MIMO data.

Model Validation: Since the aim of this closed-loop SID methodology is to substitute the open-loop SID method for re-identification, the identified models should be compared to models obtained from open-loop tests. Thus, together with the standard *validation* tests used in open-loop SID, the methodology should also be validated by comparing the closed-loop identified models to the open-loop identified models and by doing an examination of the resulting closed-loop systems. The chosen validation tests are summarised in Sections 4.8.5 and 4.8.6. It is also recommended that the models should be validated with validation data, i.e. data not used to fit the model.

The proposed methodology is summarised in terms of the five SID steps in Section 4.8.

In the following two chapters the chosen methodology is validated as well as evaluated. In the next chapter the results obtained from implementing the method on data acquired from a simulation of a 2x2 MIMO plant, controlled by an MPC controller, are discussed.