5

DATA HANDLING AND PARAMETER ESTIMATION

Authors:
Gürkan Sin
Krist V. Gernaey

Reviewer:
Sebastiaan C.F. Meijer
Juan A. Baeza

5.1 INTRODUCTION

Modelling is one of the key tools at the disposal of modern wastewater treatment professionals, researchers and engineers. It enables them to study and understand complex phenomena underlying the physical, chemical and biological performance of wastewater treatment plants at different temporal and spatial scales.

At full-scale wastewater treatment plants (WWTPs), mechanistic modelling using the ASM framework and concept (e.g. Henze *et al.,* 2000) has become an important part of the engineering toolbox for process engineers. It supports plant design, operation, optimization and control applications. Models have also been increasingly used to help take decisions on complex problems including the process/technology selection for retrofitting, as well as validation of control and optimization strategies (Gernaey *et al.,* 2014; Mauricio-Iglesias *et al.,* 2014; Vangsgaard *et al.,* 2014; Bozkurt *et al.,* 2015).

Models have also been used as an integral part of the comprehensive analysis and interpretation of data obtained from a range of experimental methods from the laboratory, as well as pilot-scale studies to characterise and study wastewater treatment plants. In this regard, models help to properly explain various kinetic parameters for different microbial groups and their activities in WWTPs by using parameter estimation techniques. Indeed, estimating parameters is an integral part of model development and application (Seber and Wild, 1989; Ljung, 1999; Dochain and Vanrolleghem, 2001; Omlin and Reichert, 1999; Brun *et al.,* 2002; Sin *et al.,* 2010) and can be broadly defined as follows:

Given a model and a set of data/measurements from the experimental setup in question, estimate all or some of the parameters of the model using an appropriate statistical method.

The focus of this chapter is to provide a set of tools and the techniques necessary to estimate the kinetic and stoichiometric parameters for wastewater treatment processes using data obtained from experimental batch activity tests. These methods and tools are mainly

intended for practical applications, i.e. by consultants, engineers, and professionals. However, it is also expected that they will be useful both for graduate teaching as well as a stepping stone for academic researchers who wish to expand their theoretical interest in the subject. For the models selected to interpret the experimental data, this chapter uses available models from literature that are mostly based on the Activated Sludge Model (ASM) framework and their appropriate extensions (Henze *et al.*, 2000).

The chapter presents an overview of the most commonly used methods in the estimation of parameters from experimental batch data, namely: (*i*) data handling and validation, (*ii*) parameter estimation: maximum likelihood estimation (MLE) and bootstrap methods, (*iii*) uncertainty analysis: linear error propagation and the Monte Carlo method, and (*iv*) sensitivity and identifiability analysis.

5.2 THEORY AND METHODS

5.2.1 Data handling and validation

5.2.1.1 Systematic data analysis for biological processes

Most activated sludge processes can be studied using simplified process stoichiometry models which rely on a 'black box' description of the cellular metabolism using measurement data of the concentrations of reactants (pollutants) and products e.g. CO_2, intermediate oxidised nitrogen species, etc. Likewise, the Activated Sludge Model (ASM) framework (Henze *et al.,* 2000) relies on a black box description of aerobic and anoxic heterotrophic activities, nitrification, hydrolysis and decay processes.

A general model formulation of the process stoichiometry describing the conversion of substrates to biomass and metabolic products is formulated below (for carbon metabolism):

$$CH_2O + Y_{SO}O_2 + Y_{SN}NH_3 \rightarrow Y_{SX}X + Y_{SC}CO_2 + Y_{SP1}P_1 \ldots + Y_{SW}H_2O \qquad \text{Eq. 5.1}$$

Equation 5.1 represents a simplification of the complex metabolic 'machinery' of cellular activity into one global relation. This simplified reaction allows the calculation of the process yields including Y_{SO} (yield of oxygen per unit substrate), Y_{SN} (yield of nitrogen per unit substrate), Y_{SX} (yield of biomass per unit substrate), Y_{SC} (yield of CO_2 per unit of substrate), Y_{SP1} (yield of intermediate product P_1 per unit of substrate), and Y_{SW} (yield of water per unit of substrate).

The coefficients of this equation are written on the basis of 1 C-mol of carbon substrate. This includes growth yield for biomass, Y_{SX}, substrate (ammonia) consumption yields, Y_{SN}, oxygen consumption yields, Y_{SO}, yield for production of CO_2, Y_{SC}, and yield for water, Y_{SW}. The biomass, X, is also written on the basis of 1 C-mol and is assumed to have a typical composition of $CH_aO_bN_c$. The biomass composition can be measured experimentally, $CH_{1.8}O_{0.5}N_{0.2}$ being a typical value. Some of the yields are also measured experimentally from the observed rates of consumption and production of components in the process as follows:

$$Y_{ji} = \frac{r_i}{r_j} = \frac{q_i}{q_j} \quad \text{and} \quad Y_{ji} = Y_{ij}^{-1} \qquad \text{Eq. 5.2}$$

Where, q_i refers to the volumetric conversion/production rate of component i, i.e. the mass of component i per unit volume of the reactor per unit time (Mass i Volume^{-1} Time^{-1}), r_i refers to the measured rate of the mass of component i per unit time per unit weight of the biomass (Mass i Time^{-1} Mass biomass^{-1}) and Y_{ji} is the yield of component i per unit of component j. In the case of biomass, x, this would refer to the specific growth rate μ:

$$\mu = r_x = \frac{q_x}{x} \qquad \text{Eq. 5.3}$$

One of the advantages of using this process stoichiometry is that it allows elemental balances for C, H, N and O to be set up and to make sure that the process stoichiometry is balanced. For the process stoichiometry given in Eq 5.1, the following elemental balance for carbon will hold, assuming all the relevant yields are measured:

$$\text{C} - \text{balance}: \quad -1 + Y_{sx} + Y_{sc} + Y_{sp_i} = 0 \qquad \text{Eq. 5.4}$$

Similarly to the carbon balance, the elemental balance for N, O and H can also be performed. Usually in biological process studies, the yield coefficient for water, Y_{SW}, is ignored because the production of water is negligible compared with the high flow rates typically treated in WWTPs. For this reason, H and O balances and

process stoichiometry are usually not closed in wastewater applications. However, the balance for the degree of reduction is closed in wastewater treatment process stoichiometry. This is the framework on which ASM is based. The degree of reduction balance is relevant since most biological reactions involve reduction-oxidation (redox)-type chemical conversion reactions in metabolism activities.

5.2.1.2 Degree of reduction analysis

A biological process will convert a substrate i.e. the input to a metabolic pathway, into a product that is in a reduced or oxidized state relative to the substrate. In order to perform redox analysis on a biological process, a method to calculate the redox potential of substrates and products is required. In the ASM framework and other biotechnological applications (Heijnen, 1999; Villadsen *et al.,* 2011), the following methodology is used:

1) Define a standard for the redox state for the balanced elements, typically C, O, N, S and P.
2) Select H_2O, CO_2, NH_3, H_2SO_4, and H_3PO_4 as the reference redox-neutral compounds for calculating the redox state for the elements O, C, N, S, and P respectively. Moreover, a unit of redox is defined as H = 1. With these definitions, the following redox levels of the five listed elements are obtained: O = -2, C = 4, N = -3, S = 6 and P = 5.
3) Calculate the redox level of the substrate and products using the standard redox levels of the elements. Several examples are provided below:
   a) Glucose ($C_6H_{12}O_6$): $6 \cdot 4 + 12 \cdot 1 + 6 \cdot (-2) = 24$. Per 1 C-mol, the redox level of glucose becomes, $\gamma_g = 24/6 = 4$ mol e^- C-mol^{-1}.
   b) Acetic acid ($C_2H_4O_2$): $2 \cdot 4 + 4 \cdot 1 + 2 \cdot (-2) = 8$. Per 1 C-mol, the redox level of Hac becomes, $\gamma_a = 8/2 = 4$ mol e^- C-mol^{-1}
   c) Propionic acid ($C_3H_6O_2$): $3 \cdot 4 + 6 \cdot 1 + 2 \cdot (-2) = 14$. Per 1 C-mol, the redox level of HPr becomes, $\gamma_p = 14/3 = 4.67$ mol e^- C-mol^{-1}.
   d) Ethanol (C_2H_6O): $2 \cdot 4 + 6 \cdot 1 + 1 \cdot (-2) = 12$. Per 1 C-mol, the redox level of HAc becomes, $\gamma_e = 12/2 = 6$ mol e^- C-mol^{-1}.
4) Perform a degree of reduction balance over a given process stoichiometry (see Example 5.1).

Example 5.1 Elemental balance and degree of reduction analysis for aerobic glucose oxidation

General process stoichiometry for the aerobic oxidation of glucose to biomass:

$$CH_2O + Y_{SO}O_2 + Y_{SN}NH_3 \rightarrow Y_{SX}X + Y_{SC}CO_2 \qquad \text{Eq. 5.5}$$

Assuming the biomass composition X is $CH_{1.8}O_{0.5}N_{0.2}$. The degree of reduction for biomass is calculated assuming the nitrogen source is ammonia (hence the nitrogen oxidation state is -3, γ_X: $4 + 1.8 + 0.5 \cdot (-2) + 0.2 \cdot (-3) = 4.2$ mol e^- C-mol^{-1}.

Now C, N and the degree of reduction balances can be performed for the process stoichiometry as follows:

Carbon balance: $-1 + Y_{SX} + Y_{SC} = 0$ Eq. 5.6

Nitrogen balance: $-Y_{SN} + 0.2 \cdot Y_{SX} = 0$ Eq. 5.7

Redox balance:

$$-1 \cdot \gamma_g - \gamma_{O2} \cdot Y_{SO} - \gamma_{NH3} \cdot Y_{SN} + \gamma_X \cdot Y_{SX} + \gamma_{CO2} \cdot Y_{SC} = 0$$
$$-1 \cdot \gamma_g - \gamma_{O2} \cdot Y_{SO} - 0 \cdot Y_{SN} + \gamma_X \cdot Y_{SX} + 0 \cdot Y_{SC} = 0$$
Eq. 5.8.

In these balance equations, there are four unknowns (Y_{SN}, Y_{SO}, Y_{SX}, Y_{SC}). Since three equations are available, only one measurement of the yield is necessary to calculate all the others. For example, in ASM applications, biomass growth yield is usually assumed measured or known, hence the other remaining yields can be calculated as follows:

CO_2 yield: $Y_{SC} = 1 - Y_{SX}$ Eq. 5.9

NH_3 yield: $Y_{SN} = 0.2 Y_{SX}$ Eq. 5.10

O_2 yield: $Y_{SO} = \frac{\gamma_g - \gamma_x \cdot Y_{SX}}{\gamma_{O2}} = \frac{4 - 4.2 Y_{SX}}{4}$ Eq. 5.11

With these coefficients known, the process stoichiometry model for 1 C-mol of glucose consumption becomes as follows:

$$CH_2O + \frac{4 - 4.2 Y_{SX}}{4} \cdot O_2 + 0.2 Y_{SX} \cdot NH_3 \rightarrow Y_{SX} \cdot X + (1 - Y_{SX}) \cdot CO_2 \qquad \text{Eq. 5.12}$$

In the ASM framework, the process stoichiometry is calculated using a unit production of biomass as a reference. Hence, the coefficients of Eq. 5.12 can be re-arranged as follows:

$$\frac{1}{Y_{SX}} \cdot CH_2O + \frac{4-4.2Y_{SX}}{4Y_{SX}} \cdot O_2 + 0.2NH_3 \rightarrow X + \left(\frac{1}{Y_{SX}} - Y_{SX}\right) \cdot CO_2 \quad \text{Eq. 5.13}$$

The unit conversion from a C-mol to a g COD basis, being the unit of ASM models, is defined using O_2 as the reference compound. Accordingly, 1 g COD is defined as -1 g O_2. From the degree of reduction of oxygen, the conversion to COD from one unit redox (mol e^-) is calculated as follows:

$$\frac{\text{Molecular weight of } O_2}{\text{Degree of reduction of } O_2} = \frac{MW_{O2}}{\gamma_{O2}} = \frac{32}{4} = 8 \text{ g COD } L^{-1} (\text{mol } e^-)^{-1} \quad \text{Eq. 5.14}$$

To convert a C-mol to a g COD basis, the unit redox needs to be multiplied with the degree of reduction of the substrate as follows:

$$\left(\frac{\text{mol } \bar{e}}{\text{C-mole}}\right) \cdot \left(\frac{\text{g COD}}{\text{mol } \bar{e}}\right) = \gamma_g \cdot 8 \frac{\text{g COD}}{\text{C-mol}} \quad \text{Eq. 5.15}$$

5.2.1.3 Consistency check of the experimental data

The value of performing elemental balances around data collected from experiments with biological processes is obvious: to confirm the data consistency with the first law of thermodynamics, which asserts that energy (in the form of matter, heat, etc.) is conserved. A primary and obvious requirement for performing elemental balances is that the model is checked and consistent. Experimental data needs to be checked for gross (measurement) errors that may be caused by incorrect calibration or malfunction of the instruments, equipment and/or sensors.

Inconsistency in the data can be checked from the sum of the elements that make up the substrates consumed in the reaction (e.g. glucose, ammonia, oxygen, etc.). This should equal to the sum of the elements (products) produced in the reaction (therefore also see Eq. 5.4 for the carbon balance). Deviation from this elemental balance indicates an incorrectly defined system description, a model inconsistency and/or measurement flaws.

In addition to the elemental balances, the degree of reduction balance provides information about whether the right compounds are included for a given pathway or whether a compound is missing in the process stoichiometry. Adding this check is helpful and provides consistency with the bioenergetic principles of biological processes (Roels, 1980; Heijnen, 1999; Villadsen *et al.,* 2011).

The consistency checks and the elemental balances (in addition to the charge balances) are included in the ASM framework as a conservation matrix to verify the internal consistency of the yield coefficients (Henze *et al.*, 2000).

The elemental composition and degree of reduction can be performed systematically using the following generic balance equation in order to test the consistency of the measured data:

$$\sum_{j=1}^{N} e_{sj} q_{sj} + e_x q_x + \sum_{j=1}^{M} e_{pj} q_{pj} = 0 \quad \text{Eq. 5.16}$$

The equation above is formulated for a biological process with N substrates and M metabolic products. In the equation, e is the elemental composition (C, H, O and N) for a component, and q the volumetric production (or consumption) rate for substrates (q_{sj}), biomass (q_x) and metabolic products (q_{pj}). Hence, the elemental balance can be formulated as follows:

$$E \cdot q = 0 \quad \text{Eq. 5.17}$$

In this equation, E is the conservation matrix and its columns refer to each conserved element and property, e.g. C, H, O, N, γ, etc. Each row of matrix E contains values of a conserved property related to substrates, products and biomass; q is a column vector including the measured volumetric rates for each compound. This is substrate as well as products and biomass.

The total number of columns in E is the number of compounds, which is the sum of substrates (N), products (M) and biomass, hence N + M + 1. The total number of constraints is 5 (C, H, O, N and γ). This means that N+M-4 is the number of degrees of freedom that needs to be measured or specified in order to calculate all the rates.

Typically, not all the rates will be measured in batch experiments. Therefore, let us assume q_m is the measured set of volumetric rates and q_u the unmeasured set of rates

which need to be calculated. In this case Eq. 5.17 can be reformulated as follows:

$$E_m q_m + E_u q_u = 0 \qquad \text{Eq. 5.18}$$
$$q_u = -\left(E_u\right)^{-1} E_m q_m = 0$$

Provided that the inverse of E_u exists ($\det(E_u) \neq 0$), Eq. 5.18 provides a calculation/estimation of the unmeasured rates in a biological process. These estimated rates are valuable on their own, but can also be used for validation purposes if redundant measurements are available. This systematic method of data consistency check is highlighted in Example 5.2.

All these calculations help to verify and validate the experimental data and measurement of the process yield. The data can now be used for further kinetic analysis and parameter estimation.

5.2.2 Parameter estimation

Here we recall a state-space model formalism to describe a system of interest. Let y be a vector of outputs resulting from a dynamic model, f, employing a parameter vector, θ; input vector, u; and state variables, x:

$$\frac{dx}{dt} = f\left(x, \theta, u, t\right); \quad x\left(0\right) = x_0 \qquad \text{Eq. 5.19}$$
$$y = g\left(x, \theta, u, t\right)$$

The above equation describes a system (a batch setup or a full WWTP) in terms of a coupled ordinary differential equations (ODE) and algebraic system of equations using a state-space formalism.

The problem statement for parameter estimation reads then as follows: for a given set of measurements, *y*, with its measurement noise collected from the system of interest, and given the model structure in Eq. 5.19, estimate the unknown model parameters (θ).

The solution approaches to this problem can be broadly classified as the manual trial and error method, and formal statistical methods.

5.2.2.1 The manual trial and error method

This approach has no formal scientific basis except for a practical motivation that has to do with getting a good model fit to the data. It works as follows: the user chooses one parameter from the parameter set and then changes it incrementally (increases or decreases around its nominal value) until a reasonable model fit is obtained to the measured data. The same process may be iterated for another parameter. The fitting process is terminated when the user deems that the model fit to data is good. This is often determined by practical and/or time constraints because this procedure will never lead to an optimal fit of the model to the measured data. In addition, multiple different sets of parameter values can be obtained which may not necessarily have a physical meaning. The success of this procedure often relies on the experience of the modeller in selecting the appropriate parameters to fit certain aspects of the measured data. Although this approach is largely subjective and suboptimal, the approach is still widely used in industry as well as in the academic/research environment. Practical data quality issues do not often allow the precise determination of parameters. Also not all (commercial) modelling software platforms provide the appropriate statistical routines for parameter estimation. There are automated procedures for model calibration using algorithms such as statistical sampling techniques, optimization algorithm, etc. (Sin *et al.,* 2008). However, such procedures focus on obtaining a good fit to experimental data and not necessarily on the identifiability and/or estimation of a parameter from a data set. This is because the latter requires proper use of statistical theory.

5.2.2.2 Formal statistical methods

In this approach, a proper statistical framework is used to suggest the problem, which is then solved mathematically by using appropriate numerical solution strategies, e.g. minimization algorithms or sampling algorithms. Under this category, the following statistical frameworks are usually employed:

a. Frequentist framework (maximum likelihood, least squares, non-linear regression, etc.).
b. Bayesian framework (Metropolis-Hasting, Markov Chain Monte Carlo (MCMC), importance sampling, etc.).
c. Pragmatic/hybrid framework (employing some elements of the two schools of thought above, e.g. the bootstrap method, Monte Carlo filtering, etc.).

The above statistical methods are among the most commonly used and recommended here as well. In particular, we focus on the frequentist and bootstrap methods as they are more fit to the intended purpose of this chapter.

Frequentist method - maximum likelihood theory

In the parameter estimation problem we usually define parameter estimators, $\hat{\theta}$, to distinguish them from the true model parameters, θ. In the context of statistical estimation, model parameters are defined as unknown and statistical methods are used to infer their true value. This difference is subtle but important to understand and to interpret the results of parameter estimation, irrespective of the methods used.

Maximum likelihood is a general method for finding estimators, $\hat{\theta}$, from a given set of measurements, y. In this approach, the model parameters θ are treated as true, fixed values, but their corresponding estimators $\hat{\theta}$ are treated as random variables. The reason is that the estimators depend on the measurements, which are assumed to be a stochastic process:

$$y = f(\theta) + \varepsilon \quad \text{where} \quad \varepsilon \propto N(0,\sigma) \qquad \text{Eq. 5.20}$$

Measurement errors, ε, are defined by a probability distribution, e.g. normal distribution, N, with zero mean and standard deviation (σ). With these assumptions, the likelihood function (L) for the parameter estimation becomes as follows (Seber and Wild, 1989):

$$L(y,\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y - f(\theta)\right)^2}{2\sigma^2} \right) \qquad \text{Eq. 5.21}$$

The most likely estimate of θ is found as those parameter values that maximize the likelihood function:

$$\hat{\theta}: \min_{\theta} L(y,\theta) \qquad \text{Eq. 5.22}$$

The solution to this problem setting (5.24) is often found by optimization algorithms such as simplex, interior point, genetic algorithms, simulated annealing, etc. The parameters obtained by calculating the maximum likelihood (Eq. 5.21) are the same as the parameters obtained by calculating the minimum cost function in Eq. 5.23.

The least squares method

This is a special case of the maximum likelihood method in which the measurements are assumed to be independent and identically distributed with white measurement errors having a known standard deviation, σ (Gaussian). The likelihood function becomes equivalent to minimizing the following cost (or objective) function, S(y,θ) (Seber and Wild, 1989):

$$S(y,\theta) = \sum \frac{\left(y - f(\theta)\right)^2}{\sigma^2} \qquad \text{Eq. 5.23}$$

Where, y stands for the measurement set, $f(\theta)$ stands for the corresponding model predictions, and Σ stands for the standard deviation of the measurement errors. The solution to the objective function (Eq. 5.24) is found by minimization algorithms (e.g. Newton's method, gradient descent, interior-point, Nelder-Mead simplex, genetic, etc.).

$$\hat{\theta}: \min_{\theta} S(y,\theta)$$
$$\frac{\partial}{\partial\theta} S(y,\theta)\Big|_{\hat{\theta}} = 0 \qquad \text{Eq. 5.24}$$

The solution to the above optimization problem provides the best estimate of the parameter values. The next step is to evaluate the quality of the parameter estimators. This step requires the estimation of the confidence interval of the parameter values and the pairwise linear correlation between the parameters.

The covariance matrix of parameter estimators

As a result of stochastic measurement, estimators have a degree of uncertainty. In the frequentist framework of thought, probability is defined in terms of the frequency of the occurrence of outcomes. Hence, in this method the uncertainty of the parameter estimators is defined by a 95 % confidence interval interpreted as the range in which 95 times out of 100 the values of the parameter estimators are likely to be located. This can be explained as if one performs the same measurement 100 times, and then performs the parameter estimation on these 100 sets and observes the following: 95 occurrences of the estimator values lie in the confidence interval, while 5 occurrences are outside this interval.

In order to estimate the confidence interval, first the covariance matrix ($\text{cov}(\hat{\theta})$), which contains complete information about the uncertainty of the parameter estimators of the estimators, needs to be estimated. One method to obtain $\text{cov}(\hat{\theta})$ is to use a linear approximation method through estimation of the Jacobian matrix (F.) of the parameter estimation problem (Seber and Wild, 1989):

$$\mathrm{cov}\left(\hat{\theta}\right) = s^2 \left(\mathrm{F.}'\cdot \mathrm{F.}\right)^{-1} \quad \text{where F.} = \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} \qquad \text{Eq. 5.25}$$

Where, s^2 is the unbiased estimation of σ^2 obtained from the residuals of the parameter estimation:

$$s^2 = \frac{S_{min}\left(y,\hat{\theta}\right)}{n-p} \qquad \text{Eq. 5.26}$$

Here, n is the total number of measurements, p is the number of estimated parameters, n-p is the degrees of freedom, $S_{min}(y,\hat{\theta})$ is the minimum objective function value and F. is the Jacobian matrix, which corresponds to the first order derivative of the model function, f, with respect to the parameter vector θ evaluated at $\theta = \hat{\theta}$.

The covariance matrix is a square matrix with (p×p) dimensions. The diagonal elements of the matrix are the variance of the parameter estimators, while the non-diagonal elements are the covariance between any pair of parameter estimators.

The 95 % confidence interval of the parameter estimators can now be approximated. Assuming a large *n*, the confidence intervals (the difference between the estimators and true parameter values), follow a student *t*-distribution, the confidence interval at 100 (1-α) % significance:

$$\hat{\theta}_{1-\alpha} = \hat{\theta} \pm t_{N-p}^{\alpha/2} \sqrt{diag\ \mathrm{cov}\left(\hat{\theta}\right)} \qquad \text{Eq. 5.27}$$

Where, $t_{N-p}{}^{\alpha/2}$ is the upper α/2 percentile of the t-distribution with N-p degrees of freedom, and *diag* cov($\hat{\theta}$) represents the diagonal elements of the covariance matrix of the parameters.

The pairwise linear correlation between the parameter estimators, R_{ij}, can be obtained by calculating a correlation matrix from unit standardization of the covariance matrix as follows:

$$R_{ij} = \frac{\mathrm{cov}\left(\theta_i,\theta_j\right)}{\sigma_{\theta_i} \times \sigma_{\theta_j}} \qquad \text{Eq. 5.28}$$

This linear correlation will range from [-1 1] and indicate whether or not he parameter estimator is uniquely identifiable (if the correlation coefficient is low) or correlated (if the correlation coefficient is high).

The bootstrap method

One of the key assumptions for using the maximum likelihood estimation (MLE) method as well as its simplified version, the nonlinear least squares method, is that the underlying distribution of errors is assumed to follow a normal (Gaussian) distribution.

In many practical applications, however, this condition is rarely satisfied. Hence, theoretically the MLE method for parameter estimation cannot be applied without compromising its assumptions, which may lead to over or underestimation of the parameter estimation errors and their covariance structure.

An alternative to this approach is the bootstrap method developed by Efron (1979), which removes the assumption that the residuals follow a normal distribution. Instead, the bootstrap method works with the actual distribution of the measurement errors, which are then propagated to the parameter estimation errors by using an appropriate Monte Carlo scheme (Figure 5.1).

The bootstrap method uses the original data set D(0) with its N data points, to generate any number of synthetic data sets $D^S(1);D^S(2);\ldots$, also with N data points. The procedure is simply to draw N data points with replacements from the set D(0). Because of the replacement, sets are obtained in which a random fraction of the original measured points, typically 1/e = 37 %, are replaced by duplicated original points. This is illustrated in Figure 5.1.

The application of the bootstrap method for parameter estimation in the field of wastewater treatment requires adjustment due to the nature of the data that is in the time series. Hence, the sampling is not performed from the original data points (which are the time series and indicate a particular trend). Instead, the sampling is performed from the residual errors and then added to the simulated model outputs (obtained by using reference parameter estimation) (Figure 5.1). This is reasonable because the measurement errors are what is assumed to be stochastic and not the main trend of the measured data points, which are caused by biological processes/mechanisms. Bearing this in mind, the theoretical background of the bootstrap method is outlined below.

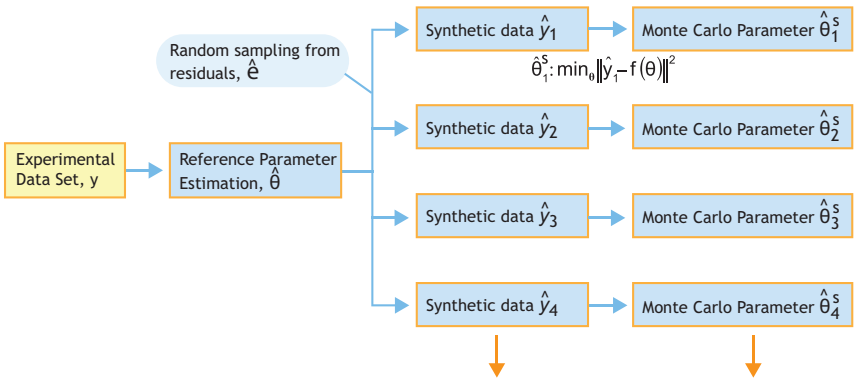

Figure 5.1 Illustration of the workflow for the bootstrap method: synthetic data sets are generated by Monte Carlo samples (random sampling with replacement) from the reference MLE. For each data set, the same estimation procedure is performed, giving M different sets of parameter estimates: $\theta^{S}_{(1)}$, $\theta^{S}_{(2)}$, … $\theta^{S}_{(M)}$.

Let us define a simple nonlinear model where y_i is the i^{th} measurement, f_i is the i^{th} model prediction, θ is a parameter vector (of length p), and ε_i is the measurement error of y_i:

$$y_i = f_i(\theta) + \varepsilon_i \quad \text{where} \quad \varepsilon_i \propto F \qquad \text{Eq. 5.29}$$

The distribution of errors, *F*, is not known. This is unlike in MLE, where the distribution is assumed *a priori*. Given y, use least squares minimization, to estimate $\hat{\theta}$:

$$\hat{\theta} : \min_{\theta} \left\| y - f(\theta) \right\|^2 \qquad \text{Eq. 5.30}$$

The bootstrap method defines $\hat{F}$ as the sample probability distribution of $\hat{\varepsilon}$ as follows:

$$\hat{F} = \frac{1}{n} \text{ (density) at } \varepsilon_i = \left(y_i - f_i(\theta) \right) \quad i = 1,2,\ldots n \qquad \text{Eq. 5.31}$$

The density is the probability of the i^{th} observation. In a uniform distribution each observation (in this case the measurement error, ε_i) has an equal probability of occurrence, where density is estimated from 1/n. The bootstrap sample, y^*, given $(\hat{\theta}, \hat{F})$, is then generated as follows:

$$y_i^* = f_i\left(\hat{\theta}\right) + \varepsilon_i^* \quad \text{where} \quad \varepsilon_i^* \propto \hat{F} \qquad \text{Eq. 5.32}$$

The realisation of measurement error in each bootstrap method, ε^*, is simulated by random sampling with replacement from the original residuals, which assigns each point with a uniform (probability) weight. By performing N random sampling with a replacement and then adding them to the model prediction (Eq. 5.31), a new synthetic data set is generated, $D^s(1) = y^*$.

By repeating the above sampling procedure M times, M data sets are generated: $D^s(1)$, $D^s(2)$, $D^s(3)$, … $D^s(M)$.

Each synthetic data set, $D^s(j)$, makes it possible to obtain a new parameter estimator $\hat{\theta}(j)$ by the same least squares minimisation method which is repeated M times:

$$\hat{\theta}_j : \min_{\theta} \left\| D^s(j) - f(\theta) \right\|^2 \quad \text{where} \quad j = 1,2\ldots M \qquad \text{Eq. 5.33}$$

The outcome from this iteration is a matrix of parameter estimators, $\hat{\theta}(M \times p)$ (M is the number of Monte Carlo samples of synthetic data and p is the number of parameters estimated). Hence, each parameter estimator now has a column vector with values. This vector of values can be plotted as a histogram and interpreted using common frequentist parameters such as the mean, standard deviation and the 95 % percentile. The covariance and correlation matrix can be computed using $\hat{\theta}(M \times p)$ itself. This effectively provides all the needed information on the quality of the parameter estimators.

For measurement errors that follow a normal distribution, both MLE and the bootstrap method will essentially provide the same results. However, if the underlying distribution of the measurements significantly deviates from a normal distribution, the bootstrap method is expected to provide a better analysis of the confidence interval of the estimators.

5.2.3 Uncertainty analysis

5.2.3.1 Linear error propagation

In linear error propagation, the covariance matrix of the parameter estimators, cov(θ), is used to propagate measurement errors to model prediction errors and to calculate standard errors and confidence intervals of the parameter estimates. Therefore, the covariance matrix of model predictions, cov(y), can be estimated using $\mathrm{cov}(\hat{\theta})$ as follows (Seber and Wild, 1989):

$$\mathrm{cov}(y) = (F.' \cdot F.)\,\mathrm{cov}(\hat{\theta})\,(F.' \cdot F.)^{-1} \qquad \text{Eq. 5.34}$$

In a similar fashion, the 1-α confidence interval of the predictions, y, can be approximated as follows:

$$y_{1-\alpha} = y \pm t_{N-p}^{\alpha/2} \sqrt{diag\,\mathrm{cov}(y)} \qquad \text{Eq. 5.35}$$

This concludes parameter estimation, confidence intervals and prediction uncertainty as viewed from the point of view of the frequentist analysis.

5.2.3.2 The Monte Carlo method

The Monte Carlo (MC) method was originally used to calculate multi-dimensional integrals and its systematic use started in the 1940s with the 'Los Alamos School' of mathematicians and physicists, namely Von Neumann, Ulam, Metropolis, Kahn, Fermi and their collaborators. The term was coined by Ulam in 1946 in honour of a relative who was keen on gambling (Metropolis and Ulam, 1949).

Within the context of uncertainty analysis, which is concerned with estimating the error propagation from a set of inputs to a set of model outputs, the integral of interest is the calculation of the mean and variance of the model outputs which are themselves indeed multidimensional integrals (the dimensionality number is determined by the length of the vector of input parameters):

$$I = \int f(x)\,dx = \int f(u_1 \ldots u_d)\,d^d x \qquad \text{Eq. 5.36}$$

Authors consider the integral of a function f(x) with x as the input vector $x = (u_1, \ldots u_d)$. Hence, the integral is taken on the d variables u_1, .., u_d over the unit hypercube $[0, 1]^d$. In the parameter estimation, these input variables are parameters of the model that have a certain range with lower and upper bounds. We assume that f is square-integrable, which means that a real value solution exists at each integration point. As a short-hand notation we will denote a point in the unit hypercube by $x = (u_1, .. u_d)$ and the function evaluated at this point by $f(x) = f(u_1, .. u_d)$, and then the multidimensional integration operation is given by:

$$E = \frac{1}{N}\sum_{1}^{N} f(x_N) \qquad \lim_{N\to\infty} \frac{1}{N}\sum_{1}^{N} f(x_N) = I \qquad \text{Eq. 5.37}$$

The law of large numbers ensures that the MC estimate (E) converges to the true value of this integral. However, as most of the time N is finite (a sampling number from input space u with dxd dimension), there will be an error in the Monte Carlo integration of multidimensional functions. This Monte Carlo integration error is scaled like $1/\sqrt{N}$. Hence, the average Monte Carlo integration error is given by $MCerr = \sigma(f)/\sqrt{N}$, where σ(f) is the standard deviation of the error, which can be approximated using sample variance:

$$\sigma^2(f) \approx s^2(f) = \frac{1}{N-1}\sum_{1}^{N} (f(x_N) - E)^2 \qquad \text{Eq. 5.38}$$

For notational simplicity, we consider the following simple model: y = f(x), where the function f represents the model under study, $x:[x_1; \ldots x_d]$ is the vector of the model inputs, and $y:[y_1; \ldots y_n]$ is the vector of the model predictions.

The goal of an uncertainty analysis is to determine the uncertainty in the elements of y that results from uncertainty in the elements of x. Given uncertainty in the vector x characterised by the distribution functions $D=[D_1, \ldots D_d]$, where D_1 is the distribution function associated with x_1, the uncertainty in y is given by:

$$\mathrm{var}(y) = \int ((y) - f(x))^2 dx$$
$$E(y) = \int f(x)\,dx \qquad \text{Eq. 5.39}$$

Where, var(y) and E(y) are the variance and expected value respectively of a vector of random variables, y, which are computed by the Monte Carlo sampling technique. In addition to the variance and mean values, one can also easily compute a percentile for y including the 95% upper and lower bounds.

5.2.4 Local sensitivity analysis and identifiability analysis

5.2.4.1 Local sensitivity analysis

Most of the sensitivity analysis results reported in the literature are of a local nature, and these are also called one factor at a time (OAT) methods. In OAT methods, each input variable is varied (also called perturbation) one at a time around its nominal value, and the resulting effect on the output is measured. The sensitivity analysis results from these methods are useful and valid in close proximity to the parameters analysed, hence the name local. In addition, the parameter sensitivity functions depend on the nominal values used in the analysis. Alternative methods, such as regional or global methods, expand the analysis from one point in the parameter space to cover a broader range in the entire parameter space but this is beyond the scope of this chapter (interested readers can consult literature elsewhere such as Saltelli *et al.*, 2000; Sin *et al.,* 2009).

The local sensitivity measure is commonly defined using the first order derivative of an output, y = f(x), with respect to an input parameter, x:

$$\text{Absolute sensitivity: } sa = \frac{\partial y}{\partial x} \qquad \text{Eq.5.40}$$

(effect on y by perturbing x around its nominal value x^0).

$$\text{Relative sensitivity: } sr = \frac{\partial y}{\partial x}\frac{x^\circ}{y^\circ} \qquad \text{Eq. 5. 41}$$

(relative effect of y by perturbing x with a fixed fraction of its nominal value x^0).

The relative sensitivity functions are non-dimensional with respect to units and are used to compare the effects of model inputs among each other.

These first-order derivatives can be computed analytically, for example using Maple or Matlab symbolic manipulation toolbox software. Alternatively, the derivatives can be obtained numerically by model simulations with a small positive or negative perturbation, Δx, of the model inputs around their nominal values, x^0. Depending on the direction of the perturbation, the sensitivity analysis can be approximated using the forward, backward or central difference methods:

Forward perturbation:

$$\frac{\partial y}{\partial x} = \frac{f\left(x^0+\Delta x\right) - f\left(x^0\right)}{\Delta x} \qquad \text{Eq. 5.42}$$

Backward perturbation:

$$\frac{\partial y}{\partial x} = \frac{f\left(x^0\right) - f\left(x^0-\Delta x\right)}{\Delta x} \qquad \text{Eq. 5.43}$$

Central difference:

$$\frac{\partial y}{\partial x} = \frac{f\left(x^0+\Delta x\right) - f\left(x^0-\Delta x\right)}{2\Delta x} \qquad \text{Eq. 5.44}$$

When an appropriately small perturbation step, Δx, is selected (usually a perturbation factor, $\varepsilon = 10^{-3}$ is used. Hence $\Delta x = \varepsilon \cdot x$), all three methods provide exactly the same results.

Once the sensitivity functions have been calculated, they can be used to assess the parameter significance when determining the model outputs. Typically, large absolute values indicate high parameter importance, while a value close to zero implies no effect of the parameter on the model output (hence the parameter is not influential). This information is useful to assess parameter identifiability issues for the design of experiments.

5.2.4.2 Identifiability analysis using the collinearity index

The first step in parameter estimation is determining which sets of parameters can be selected for estimation. This problem is the subject of identifiability analysis, which is concerned with identifying which subsets of parameters can be identified *uniquely* from a given set of measurements. Thereby, it is assumed a model can have a number of parameters. Here the term *uniquely* is

important and needs to be understood as follows: a parameter estimate is unique when its value can be estimated independently of other parameter values and with sufficiently high accuracy (i.e. a small uncertainty). This means that the correlation coefficient between any pair of parameters should be low (e.g. lower than 0.5) and the standard error of parameter estimates should be low (e.g. the relative error of the parameter estimate, σ_θ/θ, lower than e.g. 25%). As it turns out, many parameter estimation problems are ill-conditioned problems. A problem is defined ill-conditioned when the condition number of a function/matrix is very high, which is caused by multicollinearity issues. In regression problems, the condition number is used as a diagnostic tool to identify parameter identifiability issues. Such regression diagnostics are helpful in generating potential candidates of the parameter subsets for estimation which the user can select from.

There are several identifiability tests suggested in literature that are entirely based on the sensitivity functions of the parameters on the outputs. Here we are using the two-step procedure of Brun *et al*., 2002. Accordingly, the procedure works as follows: (*i*) assessment of the parameter significance ranking, (*ii*) collinearity analysis (dependency analysis of the parameter sensitivity functions in a parameter subset):

Step 1. Rank the significance of the parameters: δ^{msqr}

$$\delta^{msqr} = \sqrt{\frac{1}{N}\sum_i^N (sr_i)} \quad \text{Eq. 5.45}$$

Where, sr is a vector of non-dimensional sensitivity values, sr = i...N values.

Step 2. Calculate the collinearity index of a parameter subset K, γ_K.

$$\gamma_K = \frac{1}{\sqrt{\min\lambda_K}} \quad \text{Eq. 5.46}$$

$$\lambda_K = \text{eigen}(snorm_K^T\ snorm_K) \quad \text{Eq. 5.47}$$

$$snorm = \frac{sr}{\|sr\|} \quad \text{Eq. 5.48}$$

Where, K indicates a parameter subset, snorm is the normalized non-dimensional sensitivity function using the Euclidian norm, and λ_K represents the eigenvalues of the normalized sensitivity matrix for parameter subset K.

In Step 1, parameters that have negligible or near-zero influence on the measured model outputs are screened out from consideration for parameter estimation. In the second step, for each parameter subset (all the combinations of the parameter subsets which include 2, 3, 4,…m parameters) the collinearity index is calculated. The collinearity index is the measure of the similarity between any two vectors of the sensitivity functions. Subsets that have highly similar sensitivity functions will tend to have a very large number ($\gamma_K \sim$ inf), while independent vectors will have a smaller value $\gamma_K \sim 1$ which is desirable. In identifiability analysis, a threshold value of 5-20 is usually used in literature (Brun *et al.*, 2001; Sin and Vanrolleghem, 2007; Sin *et al.*, 2010). It is noted that this γ_K value is to be used as guidance for selecting parameter subsets as candidates for parameter estimation. The best practice is to iterate and try a number of higher ranking subsets.

5.3 METHODOLOGY AND WORKFLOW

5.3.1 Data consistency check using an elemental balance and a degree of reduction analysis

The following workflow is involved in performing an elemental balance and a degree of reduction analysis:

Step 1. Formulate a black box process stoichiometry for the biological process.

In this step, the most relevant reactants and products consumed and produced in the biological process are identified and written down. The output is a list of reactants and products for Step 2.

Step 2. Compose the elemental composition matrices (E_m and E_u).

First establish which variables of interest are measured and then define the matrices as follows: E_m includes the elemental composition and the degree of reductions for these measured variables, while E_u includes those of unmeasured variables. To calculate the degree of reduction, use the procedure given in Section 5.2.1.2.

Step 3. Compute the unmeasured rates of the species (q_u).

Using E_m and E_u together with the vector of the measured rates (q_m), the unmeasured rates (q_u) are estimated from the solution of the linear set of equations in Eq. 5.18.

Step 4. Calculate the yield coefficients.

In this step, since all the species rates of consumption/productions are now known, the yield coefficients can be calculated using Eq. 5.2 and the process stoichiometry can be written using the yield coefficient values.

Step 5. Verify the elemental balance.

In this step, a simple check is performed to verify if the elemental balance and degree of reduction balance are closed. If not, the procedure needs to be iterated by assuming a different hypothesis concerning the formation of by-products.

5.3.2 Parameter estimation workflow for the non-linear least squares method

This workflow assumes that an appropriate and consistent mathematical model is used to describe the data. Such a model confirms the elemental balance and degree of reduction analysis (see the workflow in Section 5.3.1). Usually these models are available from literature. Most of them are modified from ASM models with appropriate simplifications and/or additions reflecting the conditions of the batch experiment.

Step 1. Initialisation.

In this step, the initial conditions for the model variables are specified as well as a nominal set of parameters for the model. The initial conditions for the model are specified according to the experimental conditions (e.g. 10 mg NH_4-N added at time 0, k_La is a certain value, oxygen saturation at a given temperature is specified, etc.). An initial guess of the model parameters is taken from literature.

Step 2. Select the experimental data and a parameter subset for the parameter estimation.

In this step, the experimental data is reviewed for the parameter estimation and which parameters need to be estimated is defined. This can be done using expert judgement or, more systematically, a sensitivity and identifiability analysis (see Section 5.3.4).

Step 3. Define and solve the parameter estimation problem.

In this step, the parameter estimation problem is defined as a minimization problem and solved using optimization algorithms (e.g. *fminsearch* in Matlab)

Step 4. Estimate the uncertainty of the parameter estimators and model outputs.

In this step, calculate the covariance matrix of the parameter estimators and compute the parameter confidence intervals as well as the parameter correlation matrix. Given the covariance matrix of the parameter estimators, estimate the covariance matrix of the model outputs by linear error propagation.

Step 5. Review and analyse the results.

In this step, review the values of the parameter values, which should be within the range of parameter values obtained from the literature. In addition, inspect the confidence intervals of the parameter estimators. Very large confidence intervals imply that the parameter in question may not be estimated reliably and should be excluded from the subset.

Further, plot and review the results from the best-fit solution. Typically, the data and model predictions should fit well.

If the results (both parameter values) and the best fit solution to the data are not satisfactory, iterate as appropriate by going back to Step 1 or Step 2.

5.3.3 Parameter estimation workflow for the bootstrap method

The workflow of the bootstrap method follows on from Step 1, Step 2 and Step 3 of the non-linear least squares method.

Step 1. Perform a reference parameter estimation using the non-linear least squares method.

This step is basically an execution of steps 1, 2 and 3 of the workflow in the non-linear least squares technique. The output is a residual vector that is passed on to the next step. The residual vector is then plotted and reviewed. If the residuals follow a systematic pattern (it should be random) or contain outliers, this is a cause for concern as it may imply the bootstrap method is not suited for this application.

Step 2. Generate synthetic data by bootstrap sampling and repeat the parameter estimation.

Synthetic data is generated using Eq. 5.29-5.32 by performing bootstrap sampling (random sampling with replacement) from the residual vector and adding it to the model prediction obtained in Step 1. For each synthetic data, the parameter estimation in Step 1 is repeated and the output (that is, the values of the parameter estimators) is recorded in a matrix.

Step 3. Review and analyse the results.

In this step, the mean, standard deviation and the correlation matrix of the parameter estimators are computed from the recorded matrix data in Step 2. Moreover, the distribution function of the parameter estimators can be estimated and plotted using the vector of the parameter values that was obtained in Step 2.

As in Step 5 of the workflow in the non-linear least squares method, the results are interpreted and evaluated using knowledge from literature and process engineering.

5.3.4 Local sensitivity and identifiability analysis workflow

The workflow of this procedure starts with the assumption that a mathematical model is available and ready to be used to describe a set of experimental data.

Step 1. Initialisation.

A framework is defined for the sensitivity analysis by defining the experimental conditions (the initial conditions for the batch experiments) as well as a set of nominal values for the model analysis. The model is solved with these initial conditions and the model outputs are plotted and reviewed before performing the sensitivity analysis.

Step 2. Compute the sensitivity functions.

Define which outputs are measured and hence should be included in the sensitivity analysis. Define the experimental data points (every 1 min versus every 5 min).

Compute the sensitivity functions of the parameters on the outputs using a numerical difference, e.g. using a forward, backward or central difference. Plot, review and analyse the results.

Step 3. Rank the parameter significance.

Calculate the delta mean-square measure, δ^{msqr}, and rank the parameters according to this measure. Exclude any parameters that have zero or negligible impact on the outputs.

Step 4. Compute the collinearity index.

For all the parameter combinations (e.g. subset size 2, 3, 4….m), the collinearity index, γ_K, is calculated. Each parameter subset is ranked according to the collinearity index value.

Step 5. Review and analyse the results.

Based on the results from Step 3 and Step 4, identify a short list of candidates (parameter subsets) that are identifiable. Exclude these parameters from any parameter subset that has near-zero or negligible sensitivity on the outputs.

5.3.5 Uncertainty analysis using the Monte Carlo method and linear error propagation

The workflow for the Monte Carlo method includes the following steps:

Step 1. Input the uncertainty definition.

Identify which inputs (parameters) have uncertainty. Define a range/distribution for each uncertainty input, e.g. normal distribution, uniform distribution, etc. The output from the parameter estimators (e.g. bootstrap) can be used as input here.

Step 2. Sampling from the input space.

Define the sampling number, N, (e.g. 50, 100, etc.) and sample from the input space using an appropriate sampling technique. The most common sampling techniques are random sampling, Latin Hypercube sampling, etc. The output from this step is a sampling matrix, $X_{Nxm,}$ where N is the number of samples and m is the number of inputs.

Step 3. Perform the Monte Carlo simulations.

Perform N simulations with the model using the sampling matrix from Step 2. Record the outputs in an appropriate matrix form to be processed in the next step.

Step 4. Review and analyse the results.

Plot the outputs and review the results. Calculate the mean, standard deviation/variance, and percentiles (e.g. 95 %) for the outputs. Analyse the results within the context of parameter estimation quality and model prediction uncertainty. Iterate the analysis, if necessary, by going back to Step 1 or Step 2.

The workflow for linear error propagation:

The workflow is relatively straightforward as it is complementary to the covariance matrix of the parameter estimators and should be performed as part of the parameter estimation in the non-linear least squares method. It requires the covariance matrix of parameter estimators as well as the Jacobian matrix which are both obtained in Step 4 of the non-linear least squares methodology.

5.4 ADDITIONAL EXAMPLES

Example 5.2 Anaerobic fermentation of glucose

In this example, anaerobic fermentation of glucose to ethanol and glycerol as metabolic products is considered.

Step 1. Formulate the process stoichiometry.

Ammonia is assumed to be the nitrogen source for growth. The biomass composition is assumed to be $CH_{1.6}O_{0.5}N_{0.15}$. All the substrates are given on the basis of 1 C-mol, whereas nitrogen is on the basis of 1 N-mol. In this biological process, the substrates are CH_2O (glucose) and NH_3. The products are $CH_{1.61}O_{0.52}N_{0.15}$ (biomass), $CH_3O_{0.5}$ (ethanol), $CH_{8/3}O$ (glycerol) and CO_2. Water is excluded from the analysis, as its rate of production is not considered relevant to the process. This means that the H and O balances will not be considered either.

Step 2. Compose the elemental composition matrices (E_m and E_u).

As the process has six species (substrates + products) and three constraints (two elemental balances for C and N plus a degree of reduction balance), measurement of three rates is sufficient to estimate/infer the remaining rates.

To illustrate the concept, the measured rates are selected as the volumetric consumption rate of substrate ($-q_s$), the biomass production rate (q_x), and the glycerol production rate (q_g) hence the remaining rates for ammonia consumption as well as the production of ethanol and CO_2 need to be estimated using Eq. 5.18. In the measured rate vectors, a negative sign indicates the consumption of a species, while a positive sign indicates the production of a species.

Step 3 Compute the unmeasured rates of the species (q_u).

Recall Eq. 5.18, which is solved as follows:

$$E_m \cdot q_m + E_u \cdot q_u = 0$$

$$\begin{array}{c} \\ C \\ N \\ \gamma \end{array}\begin{array}{c} \begin{array}{ccc} S & X & Gly \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0.15 & 0 \\ 4 & 4.12 & 4.67 \end{bmatrix} \end{array} \cdot \begin{pmatrix} -q_s \\ q_x \\ q_g \end{pmatrix} + \begin{array}{c} \begin{array}{ccc} NH_3 & Eth & CO_2 \end{array} \\ \begin{bmatrix} 0 & 1.0 & 1.0 \\ 1 & 0 & 0 \\ 0 & 6 & 0 \end{bmatrix} \end{array} \cdot \begin{pmatrix} -q_n \\ q_e \\ q_c \end{pmatrix} = 0$$

$$q_u = -(E_u)^{-1} \cdot E_m \cdot q_m$$

$$\begin{pmatrix} -q_n \\ q_e \\ q_c \end{pmatrix} = -\left(\begin{bmatrix} 0 & 1.0 & 1.0 \\ 1 & 0 & 0 \\ 0 & 6 & 0 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0.15 & 0 \\ 4 & 4.12 & 4.67 \end{bmatrix} \cdot \begin{pmatrix} -q_s \\ q_x \\ q_g \end{pmatrix}$$

Solving the system of linear equations above yields the following solution where the three unmeasured rates are calculated as a function of the measured rates q_s, q_g and q_x:

$$\begin{pmatrix} -q_n \\ q_e \\ q_c \end{pmatrix} = \begin{pmatrix} -0.15q_x \\ 2q_s/3 - 467q_g/600 - 103q_x/150 \\ q_s/3 - 133q_g/600 - 47q_x/150 \end{pmatrix}$$

Step 4. Calculate the process yields.

Once the rates of all the products and substrates are estimated, one can then calculate the yield coefficients for the process by recalling Eq. 5.2 as follows:

$$Y_{sx} = \frac{q_x}{q_s} \quad \text{and} \quad Y_{sg} = \frac{q_g}{q_s}$$

$$Y_{sn} = \frac{q_n}{q_s} = \frac{-0.15q_x}{q_s} = -0.15Y_{sx}$$

$$Y_{se} = \frac{q_e}{q_s} = \frac{\left(2q_s/3 - 467q_g/600 - 103q_x/150\right)}{q_s} = \frac{2}{3} - 0.7783Y_{sg} - 0.6867Y_{sx}$$

$$Y_{sc} = \frac{q_c}{q_s} = \frac{\left(q_s/3 - 133q_g/600 - 47q_x/150\right)}{q_s} = \frac{1}{3} - 0.2217Y_{sg} - 0.3133Y_{sx}$$

With these yield coefficients estimated, the simplified process stoichiometry reads as follows:

$$0 = -CH_2O - 0.15Y_{sx} \cdot NH_3 \ldots + Y_{sx} \cdot CH_{1.61}O_{0.52}N_{0.15} + \left(\frac{2}{3} - 0.7783Y_{sg} - 0.6867Y_{sx}\right) \cdot CH_3O_{0.5} + Y_{sg} \cdot CH_{8/3}O + \left(\frac{1}{3} - 0.2217Y_{sg} - 0.3133Y_{sx}\right) \cdot CO_2$$

Step 5. Verify the elemental balance.

From the process stoichiometry, it is straightforward to verify that the elemental and degree of reduction balances are closed:

$$-1 + Y_{sx} + \left(\frac{2}{3} - 0.7783 \cdot Y_{sg} - 0.6867 \cdot Y_{sx}\right) + Y_{sg} + \left(\frac{1}{3} - 0.2217 \cdot Y_{sg} - 0.3133 \cdot Y_{sx}\right) = 0$$

The nitrogen balance:

$$-0 - 0.15 \cdot Y_{sx} + 0.15 \cdot Y_{sx} + 0 + 0 + 0 = 0$$

The degree of reduction balance:

$$-1 \cdot 4 + Y_{sx} \cdot 4.12 + \left(\frac{2}{3} - 0.7783 \cdot Y_{sg} - 0.6867 \cdot Y_{sx}\right) \cdot 6 + Y_{sg} \cdot 4.67 = 0$$

In the above example, three measured rates were assumed available as a minimum requirement to identify the system of linear equations. In practical applications, there might be two other situations: (*i*) redundant measurements: measurements of most or perhaps all of the species rates of production/consumption are available. In this case, the additional rate measurements can be used for data quality check and validation (where some of the measured rates could be used as a validation of the estimated coefficients from the balance analysis); (*ii*) a limited set of measurements: in this case too few rates measurements are available to uniquely estimate the unmeasured rates. If data from enough variables is not available, some variables should be fixed (e.g. this could be done iteratively in a search algorithm) to reduce the degrees of freedom. If not, there are infinite solutions. Further discussion of these techniques can be found elsewhere in relevant literature (Villadsen *et al.*, 2011; Meijer *et al.*, 2002; van der Heijden *et al.,* 1994).

Example 5.3 Estimate the parameters of the ammonium and nitrite oxidation processes using data from batch tests: the non-linear least squares method

Aerobic batch tests with a sludge sample from a pre-denitrification plant are performed to measure the parameters of the nitrifying bacteria, in particular the ammonium-oxidizing organisms (AOO) and nitrite-oxidising organisms (NOO). Following the recommended experimental procedure in literature (Guisasola *et al.,* 2005), two separate batch tests were performed as follows: (*i*) batch test 1 with added ammonium of 20 mg NH_4-N L^{-1} and an inhibitor (sodium azide) to suppress NOO activity, (*ii*) batch test 2 with added ammonium of 20 mg NH_4-N L^{-1} without any inhibitor addition. In both tests, both the pH and temperature are controlled at 7.5 and 25 °C respectively. During both the batch tests, ammonium, nitrite and nitrate are measured every 5 minutes, while dissolved oxygen is measured every minute. The data collected is shown in Figure 5.2. For the sake of simplicity and to keep the focus on the demonstration of the methods and their proper interpretation, these examples use synthetic data with random (white-noise) addition.

Part 1. Estimate the parameters of the ammonium oxidation process.

Several models are suggested in literature reviewed in Sin *et al.,* 2008. We use the following mathematical model given in Table 5.1 to describe the kinetics of ammonium and nitrite oxidation. For the sake of simplicity, the following is assumed: (*i*) endogenous

respiration related to heterotrophic biomass is constant (hence not modelled), (*ii*) the inert fraction of the biomass released during decay is negligible (hence not modelled), and (*iii*) the ammonium consumed for autotrophic growth of biomass is negligible. It is noted that for the sake of completeness all of the above phenomena should be described which makes the analysis more accurate. However, here the model is kept simple to focus the attention of the reader on the workflow of parameter estimation.

Table 5.1 The two-step nitrification model structure using matrix representation (adopted from Sin *et al.*, 2008)

Variables→C_i	S_{NH}	S_O	S_{NO2}	S_{NO3}	X_{AOO}	X_{NOO}	Rates
Processes ↓j	mg N L^{-1}	mg O_2 L^{-1}	mg N L^{-1}	mg N L^{-1}	mg COD L^{-1}	mg COD L^{-1}	q_j
AOO growth	$\frac{-1}{Y_{AOO}}$	$1-\frac{3.43}{Y_{AOO}}$	$\frac{1}{Y_{AOO}}$		1		$\mu_{max}^{AOO} \cdot M_{NH} \cdot M_{O,AOO} \cdot X_{AOO}$
AOO decay		1			−1		$b_{AOO} \cdot X_{AOO}$
NOO growth		$1-\frac{1.14}{Y_{NOO}}$	$\frac{-1}{Y_{NOO}}$	$\frac{1}{Y_{NOO}}$		1	$\mu_{max}^{NOO} \cdot M_{NO2} \cdot M_{O,NOO} \cdot X_{NOO}$
NOO decay		1				−1	$b_{NOO} \cdot X_{NOO}$
Aeration		1					$kLa \cdot (S_o^{sat} - S_o)$

$$M_{NH}: \frac{S_{NH}}{S_{NH}+K_{s,AOO}}; M_{O,AOO}: \frac{S_O}{S_O+K_{o,AOO}}; M_{O,NOO}: \frac{S_O}{S_O+K_{o,NOO}}; M_{NO2}: \frac{S_{NO2}}{S_{NO2}+K_{s,NOO}}$$

The model has in total six ordinary differential equations (ODE), which corresponds to one mass balance for each variable of interest. Using a matrix notation, each ODE can be formulated as follows:

$$\frac{dC_i}{dt} = \sum_j v_{ij} \cdot q_j \qquad \text{Eq. 5.49}$$

The model is implemented in Matlab and solved using a standard differential equation solver (ODE45 in Matlab).

```
%% solve the ODE model:
%[time,output] = ODEsolver('Model',[starttime
simulation  endtime simulation],Initial conditions for
model variables,simulation options,model parameters);
options=odeset('RelTol',1e-7,'AbsTol',1e-8);
[t,y] = ode45(@nitmod,t,x0,options,par);
```

Step 1. Initialisation.

The model has in total 12 parameters. The nominal values as well as their range are taken from literature (Sin *et al.,* 2008) and shown in Table 5.2

The model has six state variables, all of which need to be specified to solve the system of the ODE equations. The initial condition corresponding to batch test 1 is shown in Table 5.3.

Step 2. Select the measurements and parameter subset for parameter estimation.

We used data collected from batch test 1 which includes ammonium, nitrite and dissolved oxygen measurements. Due to the suppression of NOO activity, no nitrate production is observed. Since batch test 1 is not designed for decay rate coefficient estimation, we consider all the parameters of AOO except b_{AOO} for estimation. Hence, the following is our selection:

- Y = [NH_4 NO_2 DO]; selected measurement set, Y.
- θ = [Y_{AOO} μ_{max}^{AOO} $K_{s,AOO}$ $K_{o,AOO}$]; parameter subset for the estimation.

Step 3. Solve the parameter estimation problem.

The parameter estimation is programmed as a minimization problem using the sum of the squared errors as the cost function and solved using an unconstrained non-linear optimisation solver

(*fminsearch* algorithm in Matlab) using the initial parameter guess given in Table 5.2 and initial conditions in Table 5.3. To simulate the inhibitor addition, the maximum growth rate of NOO is assumed to be zero in the model simulations. The best estimates of the parameter estimators are given in Table 5.3.

Table 5.2 Nominal values of the model parameters used as an initial guess for parameter estimation together with their upper and lower bounds.

Parameter	Symbol	Unit	Nominal value	Range
Ammonium-oxidising organisms (AOO)				
Biomass yield	Y_{AOO}	mg COD mg N^{-1}	0.15	0.11 - 0.21
Maximum growth rate	μ_{max}^{AOO}	d^{-1}	0.8	0.50 - 2.10
Substrate (NH_4) affinity	$K_{s,AOO}$	mg N L^{-1}	0.4	0.14 - 1.00
Oxygen affinity	$K_{o,AOO}$	mg O_2 L^{-1}	0.5	0.10 - 1.45
Decay rate coefficient	b_{AOO}	d^{-1}	0.1	0.07 - 0.30
Nitrite-oxidising organisms (NOO)				
Biomass yield	Y_{NOO}	mg COD mg N^{-1}	0.05	0.03 - 0.09
Maximum growth rate	μ_{max}^{NOO}	d^{-1}	0.5	0.40 - 1.05
Substrate (NO_2) affinity	$K_{s,NOO}$	mg N^{-1}	1.5	0.10 - 3.00
Oxygen affinity	$K_{o,NOO}$	mg O_2 L^{-1}	1.45	0.30 - 1.50
Decay rate coefficient	b_{NOO}	d^{-1}	0.12	0.08 - 0.20
Experimental setup				
Oxygen mass transfer	k_La	d^{-1}	360	*
Oxygen saturation	$S_{O,sat}$	mg O_2 L^{-1}	8	*

* Not estimated in this example but assumed known.

Table 5.3 Initial condition of the state variables for the model in batch test 1.

Variable	Symbol	Unit	Initial value	Comment
Ammonium	S_{NH}	mg N L^{-1}	20	Pulse addition
Oxygen	S_O	mg O_2 L^{-1}	8	Saturation
Nitrite	S_{NO2}	mg N L^{-1}	0	Post denitrified
Nitrate	S_{NO3}	mg N L^{-1}	0	Post denitrified
AOO biomass	X_{AOO}	mg COD L^{-1}	75	Ratio of AOO to NOO reflects ratio of their yields
NOO biomass	X_{NOO}	m COD L^{-1}	25	

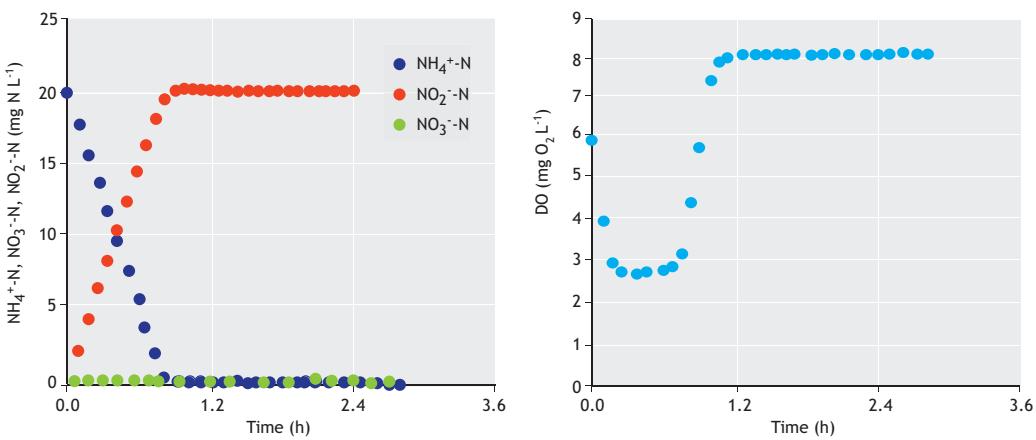

Figure 5.2 Data collected in batch test 1. NH_4, NO_2 and DO are used as the measured data set.

```
%%step 3 define and solve parameter estimation
problem (as a minimization problem)
options =optimset('display',
'iter','tolfun',1.0e-06, 'tolx',1.0e-5,
'maxfunevals', 1000);
[pmin,sse]=fminsearch(@costf,pinit,options,td,yd
,idx,iy);
```

Step 4. Estimate the uncertainty of the parameter estimators and the model prediction uncertainty.

In this step, the covariance matrix of the parameter estimators is computed. From the covariance matrix, the standard deviation, 95 % confidence interval as well as the correlation matrix are obtained. The results are shown in Table 5.4.

Table 5.4 Optimal values of the parameter estimators after the solution of the parameter estimation problem.

Parameter	Initial guess, θ°	Optimal values, $\hat{\theta}$
Y_{A00}	0.1	0.15
μ_{max}^{A00}	0.8	1.45
$K_{s,A00}$	0.4	0.50
$K_{o,A00}$	0.5	0.69

```
%% get the Jacobian matrix. use built-in
"lsqnonlin.m" but with no iteration.
options =optimset('display',
'iter','tolfun',1.0e-06, 'tolx',1.0e-5,
'maxfunevals', 0);
[~,~,residual,~,~,~,jacobian]=lsqnonlin(@costl,p
min,[],[],options,td,yd,idx,iy);
j(:,:)=jacobian; e=residual;
s=e'*e/dof; %variance of errors
%% calculate the covariance of parameter
estimators
pcov = s*inv(j'*j) ; %covariance of parameters
psigma=sqrt(diag(pcov))'; % standard deviation
parameters
pcor = pcov ./ [psigma'*psigma]; % correlation
matrix
alfa=0.025; % significance level
tcr=tinv((1-alfa),dof); % critical t-dist value
at alfa
p95 =[pmin-psigma*tcr; pmin+psigma*tcr]; %+-95%
confidence intervals
```

Table 5.5 Parameter estimation quality for the ammonium oxidation process: standard deviation, 95% confidence intervals and correlation matrix.

Parameter	Optimal value, $\hat{\theta}$	Standard deviation, σ_{θ}	95 % confidence interval (CI)		Correlation matrix			
					Y_{A00}	μ_{max}^{A00}	$K_{s,A00}$	$K_{o,A00}$
Y_{A00}	0.15	0.0076	0.130	0.160	1	0.96	0.0520	0.17
μ_{max}^{A00}	1.45	0.0810	1.290	1.610		1	0.0083	0.42
$K_{s,A00}$	0.50	0.0180	0.470	0.540			1	-0.26
$K_{o,A00}$	0.69	0.0590	0.570	0.800				1

Using the covariance matrix of the parameter estimators, the uncertainty in the model prediction is also calculated and the results are shown in Figure 5.3.

```
%% calculate confidence intervals on the model
output
ycov = j * pcov * j';
ysigma=sqrt(diag(ycov)); % std of model outputs
ys=reshape(ysigma,n,m);
y95 = [y(:,iy) - ys*tcr y(:,iy)+ys*tcr]; % 95%
confidence intervals
```

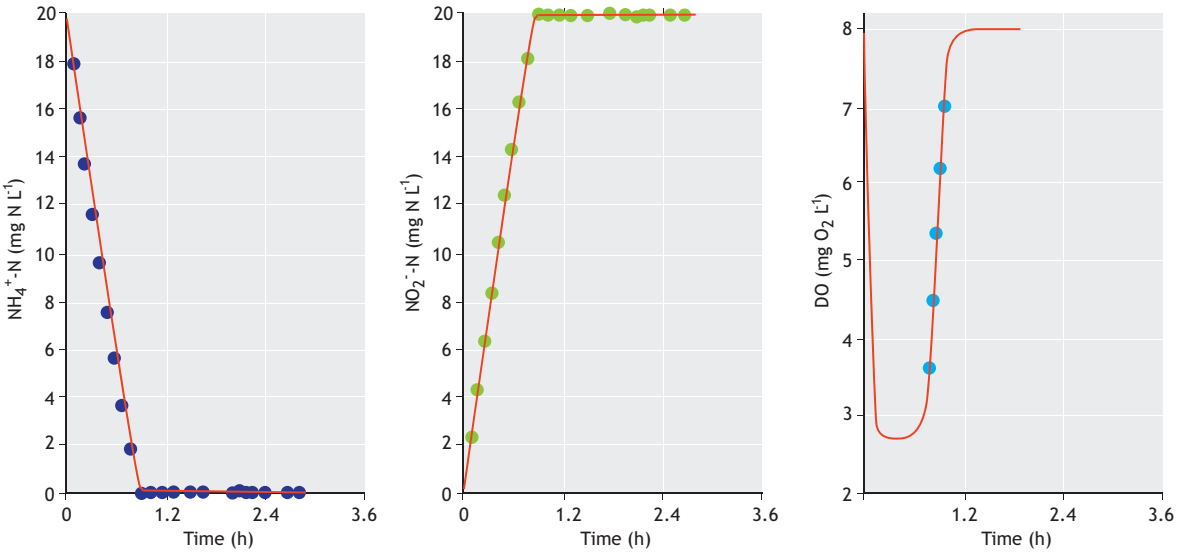

Figure 5.3 Model outputs including 95 % confidence intervals calculated using linear error propagation (red lines). The results are compared with the experimental data set.

Step 5. Review and analyse the results.

The estimated parameter values (Table 5.5) are found to be within the range reported in literature. This is an indication that the parameter values are credible. The uncertainty of these parameter estimators is found to be quite low. For example, the relative error (e.g. standard deviation/mean value of parameter values) is less than 10 %, which is also reflected in the small confidence interval. This indicates that the parameter estimation quality is good. It is usually noted that relative error higher than 50 % is indicative of bad estimation quality, while relative error below 10 % is good.

Regarding the correlation matrix, typically from estimating parameters from batch data for Monod-like models, the growth yield is significantly correlated with the maximum growth rate (the linear correlation coefficient is 0.96). Also notable is the correlation between the maximum growth rate and the oxygen affinity constant. This means that a unique estimation of the yield and maximum growth rate is not possible. Further investigation of the correlation requires a sensitivity analysis, which is demonstrated in Example 5.5.

Since the parameter estimation uncertainty is low, the uncertainty in the model predictions is also observed to be small. In Figure 5.3, the mean (or average) model prediction and the 95 % upper and lower bounds are quite close to each other. This means that the model prediction uncertainty due to parameter estimation uncertainty is negligible. It is noted that a comprehensive uncertainty analysis of the model predictions will require analysis of all the other sources of uncertainty including other model parameters as well as the initial conditions. However, this is outside the scope of this example and can be seen elsewhere (Sin *et al.,* 2010). Measurement error uncertainty is considered in Example 5.6.

This concludes the analysis of parameter estimation using the non-linear least squares method for the AOO parameters.

Part 2. Estimate the parameters for the NOO step.

Steps 1 and 2. Initial conditions and selection of data and parameter subsets for the parameter estimation.

The same initial condition for batch test 1 is used in batch test 2 but without any inhibitor addition, meaning that in this example the nitration is active. The data collected from batch test 2 is shown in Figure 5.3, which includes ammonium, nitrite, nitrate and DO measurements.

- $Y_2 = [NH_4\ NO_2\ NO_3\ DO]$; selected measurement set, Y.

The parameter values of AOO were set to the estimated values (Table 5.4) in the first part and are hence known, while the yield and kinetic parameters of NOO can be identified from the data:

- $\theta_2 = [Y_{NOO}\ \mu_{max}^{NOO}\ K_{s,NOO}\ K_{o,NOO}]$; parameter subset for the estimation.

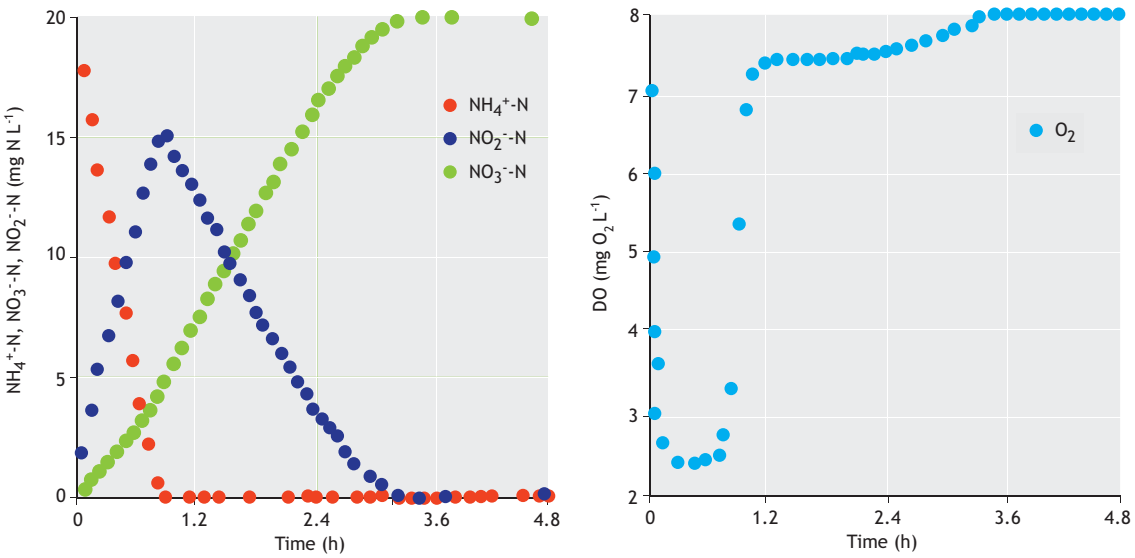

Figure 5.4 Measured data from batch test 2.

Steps 3 and 4. Solve the parameter estimation problem and calculate the parameter estimation uncertainties.

The AOO parameters are previously estimated in Part 1. The results of the solution of the parameter estimation problem as well as the parameter uncertainties for NOO are shown in Table 5.6.

.

Table 5.6 Optimal values of the parameter estimators after solution of the parameter estimation problem.

Parameter	Optimal values, $\hat{\theta}$	Standard deviation, σ_θ	95 % confidence interval (CI)		Correlation matrix			
			Lower bound	Upper bound	Y_{NOO}	μ_{max}^{NOO}	$K_{s,NOO}$	$K_{o,NOO}$
Y_{NOO}	0.04	0.01	0.01	0.07	1.00	1.00	0.54	-0.86
μ_{max}^{NOO}	0.41	0.13	0.15	0.66		1.00	0.55	-0.86
$K_{s,NOO}$	1.48	0.03	1.42	1.55			1.00	-0.37
$K_{o,NOO}$	1.50	0.05	1.39	1.60				1.00

The linear propagation of the parameter estimation error (covariance matrix) to the model prediction uncertainty is shown in Figure 5.5.

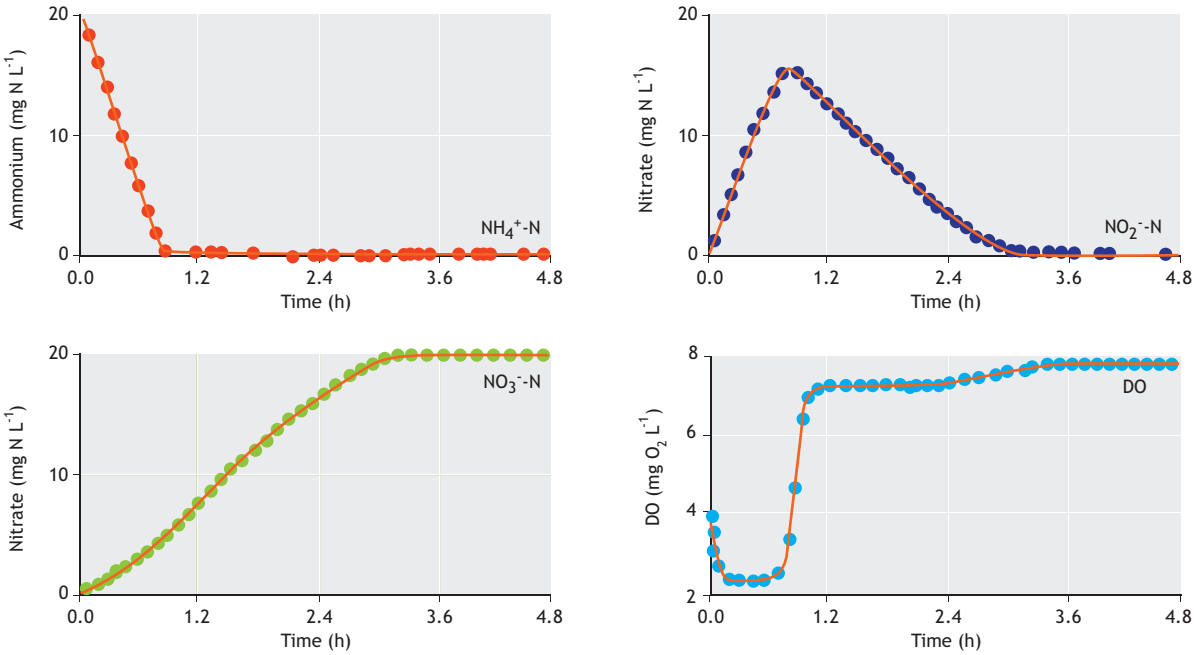

Figure 5.5 Model outputs including 95 % confidence intervals compared with the experimental data set.

Step 5. Review and analyse the results

The estimated parameter values are within the range reported for the NOO parameters in literature, which makes them credible. However, this time the parameter estimation error is noticeably higher, e.g. the relative error (the ratio of standard deviation to the optimal parameter value) is more than 30%, especially for the yield and maximum growth rate. This is not surprising since the estimation of both the yield and maximum growth rate is fully correlated (the pairwise linear correlation coefficient is 1). These statistics mean that a unique parameter estimation for the yield, maximum growth rate and oxygen half-saturation coefficient of NOO (the pairwise linear correlation coefficient is 0.86) is not possible with this batch experiment. Hence, this parameter subset should be considered as a subset that provides a good fit to the experimental data, while individually each parameter value may not have sensible/physical meaning.

The propagation of the parameter covariance matrix to the model prediction uncertainty indicates low uncertainty on the model outputs. This means that although parameters themselves are not uniquely identifiable, they can still be used to perform model predictions, e.g. to describe batch test data. While performing simulations with the model, however, one needs to report the 95% confidence intervals of the simulated values as well. The latter reflects how the covariance of the parameter estimates (implying the parameter estimation quality) affects the model prediction quality. For example, if the 95% confidence interval of the model predictions is low, then the effect of the parameter estimation error is negligible.

Part 1 and Part 2 conclude the parameter estimation for the two-step nitrification step. The results show that the quality of the parameter estimation for AOO is relatively higher than that of NOO using batch data for these experiments. This poor identifiability will be investigated later on, using sensitivity analysis to improve the identifiability of individual parameters of the model.

Regarding the model prediction errors, the 95% confidence interval of the model outputs is quite low. This means that the effects of the parameter estimation errors on the model outputs are low.

Example 5.4 Estimate the parameters of ammonium oxidation using data from the batch test – the bootstrap method

In this example, we investigate the parameter estimation problem in part 1 of Example 5.3. We used the data from batch test 1 to estimate the parameters of AOO.

Step 1. Perform a reference parameter estimation using non-linear least squares.

The workflow in this step is exactly the same as the steps 1, 2 and 3 in Example 5.3. The output from this step is the best fit to the data and the distribution of residuals (Figure 5.6).

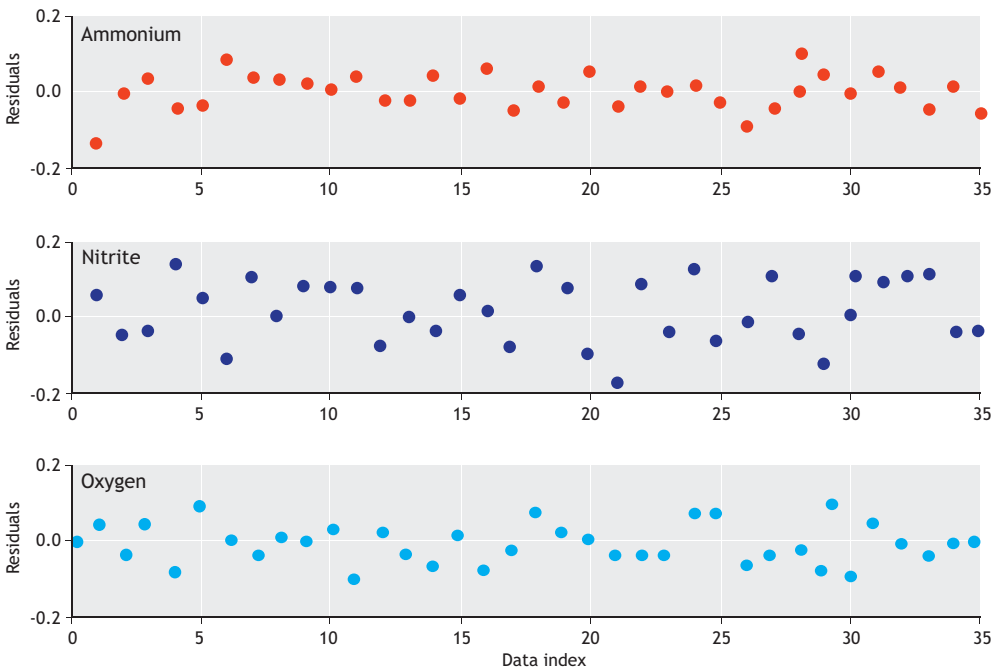

Figure 5.6 Residuals from the reference parameter estimation.

Step 2. Generate synthetic data by bootstrap sampling and repeat the parameter estimation.

In this step, bootstrap sampling from residuals is performed.

```
nboot=50; % bootstrap samples
for i=1:nboot
  disp(['the iteration number is :
',num2str(i)])
  onesam =ceil(n*rand(n,m)); % random sampling
with replacement
  rsam  =res(onesam); % measurement errors for
each variable
  ybt = y(:,iy) + rsam ; % synthetic data: error
+ model (ref PE)
  options
=optimset('display','iter','tolfun',1.0e-
06,'tolx',1.0e-5,'maxfunevals',1000);
[pmin(i,:),sse(i,:)]=lsqnonlin(@costl,pmin1,plo,
phi,options,td,ybt,idx,iy);
  bootsam(:,:,i)=ybt; % record samples
end
```

Fifty bootstrap samples from residuals (random sampling with replacement) are performed and added to the model, thereby yielding the 50 synthetic measurement data sets shown in Figure 5.7.

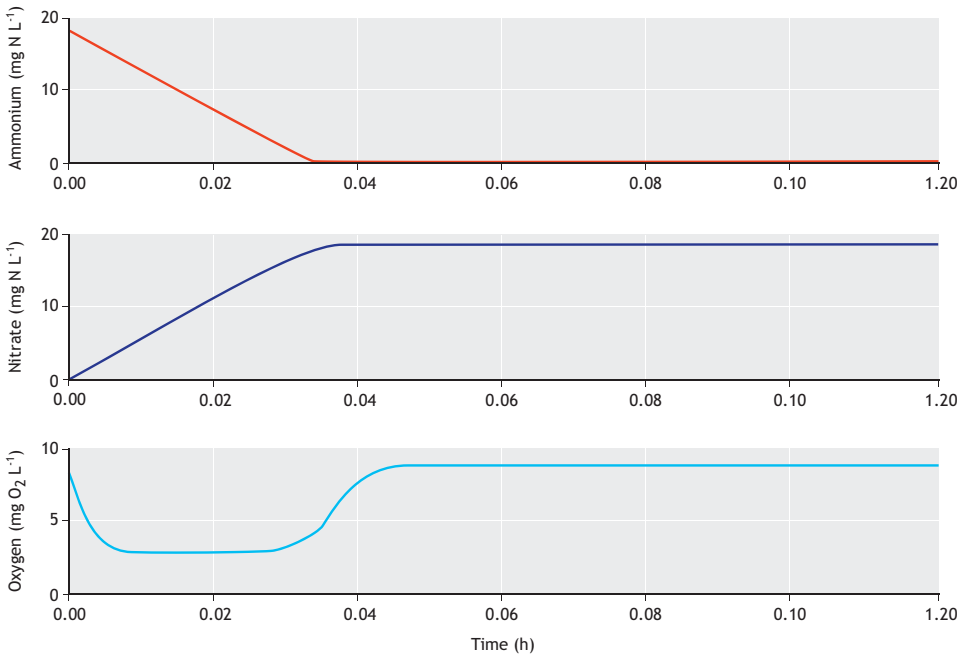

Figure 5.7 Generation of synthetic data using bootstrap sampling from the residuals (50 samples in total).

For each of this synthetic data (a bootstrap sample), a parameter estimation is performed and the results are recorded for analysis. Because 50 synthetic data sets are generated, this means that 50 different estimates of parameters are obtained. The results are shown as a histogram for each parameter estimate in Figure 5.8.

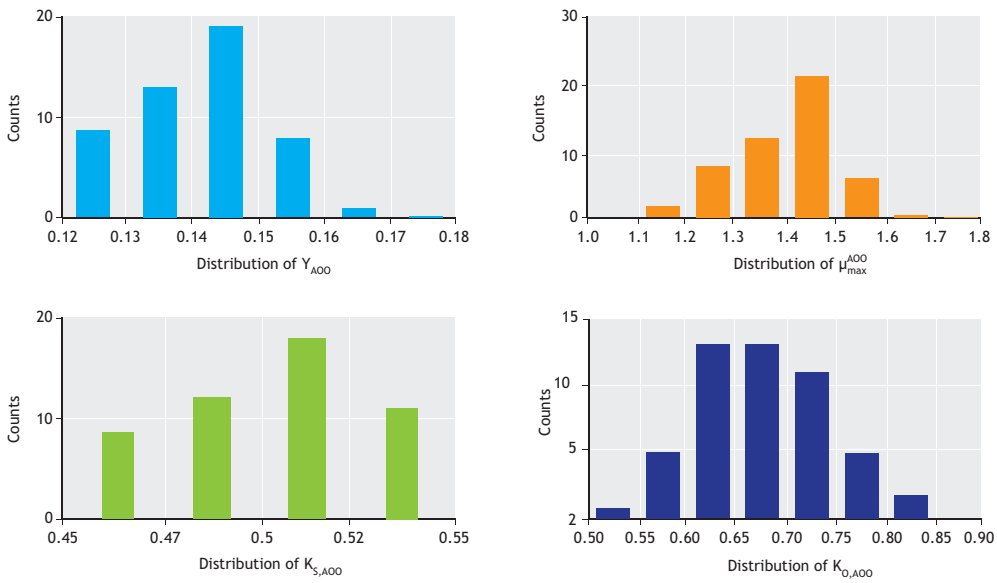

Figure 5.8 Distribution of the parameter estimates obtained using the bootstrap method (each distribution contains 50 estimated values for each parameter).

Step 3. Review and analyse the results.

Step 2 provided a matrix of the parameter estimates, θ_{50x4}. In this step, the mean, standard deviation and correlation matrix properties of this matrix are evaluated. The results are shown in Table 5.7.

```
.%step 3 Evaluate/interpret distribution of theta
disp('The mean of distribution of theta are')
disp(mean(pmin))
disp('The std.dev. of distribution of theta
are')
disp(std(pmin))
disp('')
disp('The correlation of parameters')
disp(corr(pmin))
```

Table 5.7 Optimal values of the parameter estimators after solving the parameter estimation problem.

Parameter	Optimal value, $\hat{\theta}$	Standard deviation, σ_θ	Correlation matrix			
			Y_{A00}	μ_{max}^{A00}	$K_{s,A00}$	$K_{o,A00}$
Y_{A00}	0.14	0.01	1.00	0.97	-0.03	0.20
μ_{max}^{A00}	1.40	0.11		1.00	-0.07	0.41
$K_{s,A00}$	0.50	0.02			1.00	-0.28
$K_{o,A00}$	0.68	0.07				1.00

All the results, including the mean parameter estimates, their standard deviation and the correlation matrix are in good agreement with the parameter estimates obtained from the non-linear least squares method (compare with Table 5.6.). This is expected, since the distribution of residuals is found to be quite similar to a normal distribution (Figure 5.6). In this case, both the non-linear least squares (and the linear approximation of covariance matrix estimation) as well as the bootstrap method will obtain statistically similar results.

Also the model simulation with the mean values obtained from the bootstrap samples provided similarly good fit to the measured data, as shown in Figure 5.3.

Because the bootstrap method is intuitively simple and straightforward and does not require a calculation of the Jacobian matrix, we recommend it for practical use. However, a reservation on using this method is that the distribution of the residuals should be inspected and should not contain any systematic pattern (indicating model structure or systematic measurement issues).

```
%% get the Jacobian matrix. use built-in
"lsqnonlin.m" but with no iteration.
options =optimset('display',
'iter','tolfun',1.0e-06, 'tolx',1.0e-5,
'maxfunevals', 0);
[~,~,residual,~,~,~,jacobian]=lsqnonlin(@costl,p
min,[],[],options,td,yd,idx,iy);
j(:,:)=jacobian; e=residual;
s=e'*e/dof; %variance of errors
%% calculate the covariance of parameter
estimators
pcov = s*inv(j'*j) ; %covariance of parameters
psigma=sqrt(diag(pcov))'; % standard deviation
parameters
pcor = pcov ./ [psigma'*psigma]; % correlation
matrix
alfa=0.025; % significance level
tcr=tinv((1-alfa),dof); % critical t-dist value
at alfa
p95 =[pmin-psigma*tcr; pmin+psigma*tcr]; %+-95%
confidence intervals
```

Example 5.5 Sensitivity and identifiability analysis of the ammonium oxidation process parameters in batch tests

Here the ammonium oxidation process is used as described in Example 5.3. The objective of this example is twofold: in the first part, we wish to assess the sensitivity of all the AOO parameters to all the model outputs under the experimental conditions of batch test 1. In the second part we wish to examine, given the measured data set, which parameter subsets are potentially identifiable and compare them with the parameter subset already used in the parameter estimation in Example 5.3.

Step 1. Initialisation. We use the initial conditions of batch test 1 as described in Table 5.3 as well as the nominal values of AOO model parameters as given in Table 5.2.

The model outputs of interest are:

- $y = [NH_4\ NO_2\ NO_3\ DO\ AOO\ NOO]$

The parameter set of interest is:

- $\theta = [Y_{AOO}\ \mu_{max}^{AOO}\ K_{s,AOO}\ K_{o,AOO}\ b_{AOO}]$

Step 2. Compute and analyse the sensitivity functions.

In this step, the absolute sensitivity functions are computed using numerical differentiation and the results are recorded for analysis.

```
for i=1:m; %for each parameter
  dp(i) = pert(i) * abs(ps(i)); % parameter
perturbation
  p(i)  = ps(i) + dp(i);    % forward
perturbation
  [t1,y1] = ode45(@nitmod,td,x0,options,p);
  p(i) = ps(i) - dp(i); %backward perturbation
  [t2,y2] = ode45(@nitmod,td,x0,options,p);
  dydpc(:,:,i) = (y1-y2) ./ (2 * dp(i));
%central difference
  dydpf(:,:,i) = (y1-y) ./ dp(i); %forward
difference
  dydpb(:,:,i) = (y-y2) ./ dp(i); %backward
difference
  p(i)=ps(i); % reset parameter to its reference
value
end
```

The output sensitivity functions (absolute) are plotted in Figure 5.9 for one parameter, namely the yield of AOO growth for the purpose of detailed examination. The interpretation of a sensitivity function is as follows: (*i*) higher magnitude (positive or negative alike) means higher influence, while lower or near zero magnitude means negligible/zero influence of the parameter on the output, (*ii*) negative sensitivity means that an increase in a parameter value would decrease the model output, and (*iii*) positive sensitivity means that an increase in a parameter value would increase the model output. With this in mind, it is noted that the yield of AOO has a positive effect on ammonium and an equally negative impact on nitrite. This is expected from the model structure where there is an inverse relationship between the yield and ammonium (substrate) consumption. A higher yield means less ammonium is consumed per unit growth of biomass, and hence it would also mean more ammonium present in the batch test. Since less ammonium is consumed, less nitrite would be produced (hence the negative correlation).

On the other hand, it is also noted that the sensitivity of the yield parameter increases gradually during the linear growth phase and starts to decrease as we are nearer to the depletion of ammonium. Once the ammonium is depleted, the sensitivity becomes nil as expected. As predicted, the yield has a positive impact on AOO growth since a higher yield means higher biomass production. Regarding oxygen, the yield first has a positive impact that becomes negative towards the completion of ammonium. This means there is a rather non-linear relationship between the oxygen profile and the yield parameter. As expected, the yield of AOO has no impact on the nitrate and NOO outputs in batch test 1, because of the addition of the inhibitor that effectively suppressed the second step of nitrification.

In the sensitivity analysis, what is informative is to compare the sensitivity functions among each other. This is done in Figure 5.10 using non-dimensional sensitivity functions, which are obtained by scaling the absolute sensitivity function with their respective nominal values of parameters and outputs (Eq. 5.41). Figure 5.10 plots the sensitivity of all the model parameters with respect to the six model outputs. Each subplot in the figure presents the sensitivity functions of all the parameters with respect to one model output shown in the legend. The *y*-axis indicates the non-dimensional sensitivity measure, while the *x*-axis indicates the time during the batch activity. For example, we observe that the sensitivity of parameters to nitrate and NOO is zero. This is logical since NOO activity is assumed to be zero in this simulation.

For the model outputs for ammonium, nitrite and oxygen, the sensitivity functions of the yield and maximum growth rate for AOO follow an inversely proportional trend/pattern. This inversely proportional relation is the reason why the parameter estimation problem is an ill-conditioned problem. This means that if the search algorithm increases the yield and yet at the same time decreases the maximum growth rate with a certain fraction, the effect on the model output could be cancelled out. The result is that many combinations of parameter values for the yield and maximum growth rate can have a similar effect on the model output. This is the

reason why a high correlation coefficient is obtained after the parameter estimation has been performed. This means that for a parameter to be uniquely identifiable, their sensitivity functions should be unique and not correlated with the sensitivity function of the other parameters.

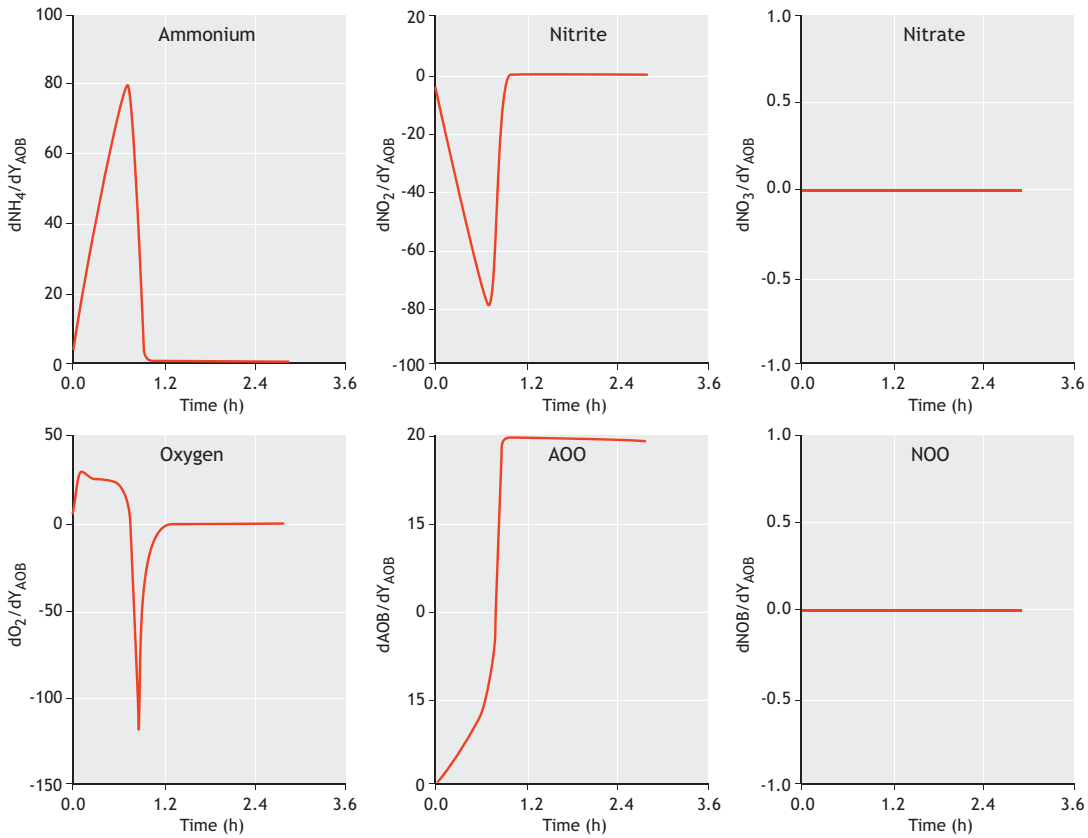

Figure 5.9 Absolute sensitivity of the AOO yield on all the model outputs.

Another point of interest regarding these plots is that the relative effect (that is, the magnitude of values on the *y*-axis) of the parameters on ammonium, oxygen and nitrite is quite similar. This means that all three of these variables are equally relevant and important for estimating these parameters.

Step 3. Parameter-significance ranking.

In this step the significance of parameters is ranked by summarizing the non-dimensional sensitivity functions of the parameters to model outputs using the δ^{msqr} measure. The results are shown in Figure 5.11.

The results show that the decay rate of AOO has almost zero effect on all three of the measured variables (ammonium, nitrite and oxygen) and therefore cannot be estimated. This is known from process engineering and for this reason, short-term batch tests are not used to determine decay constants. This result therefore is a confirmation of the correctness of the sensitivity analysis. With regards to the maximum growth rate and yield, these parameters are equally important followed by the affinity constant for oxygen and ammonium. This indicates that at least four parameters can potentially be estimated from the data set.

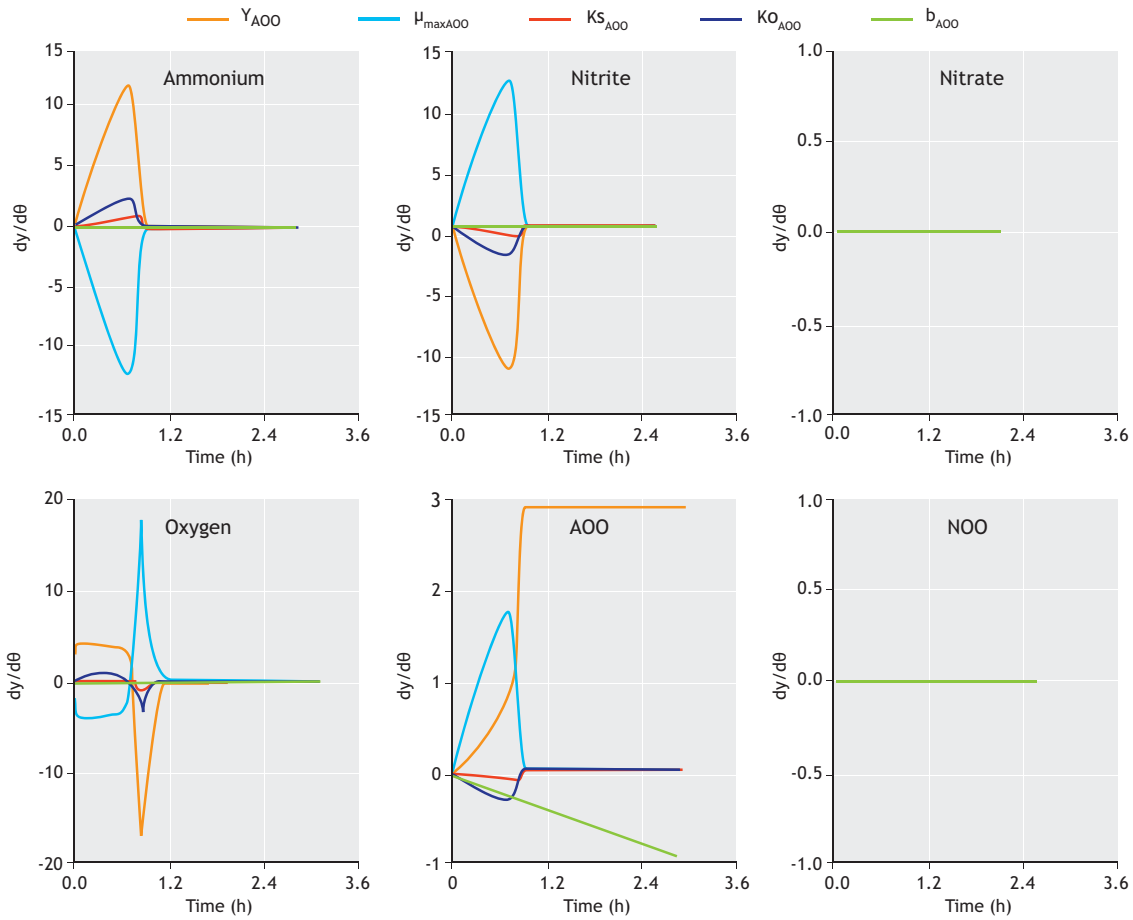

Figure 5.10 Relative sensitivity functions of the AOO parameters on the model outputs.

Step 5. Identifiability analysis.

In this step, normalized sensitivity functions are used to assess which parameter subsets have a small collinearity index. The collinearity index is a measure of how two sensitivity functions are aligned together, therefore implying linear dependency.

```
for i = 2:subset
    combos = combnk(set,i); % all possible
parameter combinations of different subset size
(2,3,4...)
    for j=1:n
        tempn   = snormy(:,combos(j,:)) ;
        tempa   = say(:,combos(j,:)) ;
        nsm     = tempn'*tempn;  % normalized
sensitivity matrix
        asm     = tempa'*tempa;   % absolute
sensitivity matrix, fim
        dtm     = sqrt(det(asm))^(1/(i*2));
%determinant index
        col     = 1/sqrt(min(eig(nsm))); %
collinearity index
        subs(j,:)    = [k i col dtm] ;
    end
end
```

The identifiability analysis indicates that there are 26 different combinations of the parameter subsets that can potentially be used for parameter estimation using the ammonium, nitrite and oxygen measurements (Table 5.8). The collinearity index value was changed from 1.2 to 53 and in general tends to increase for larger parameter subset sizes. The parameter subset K#21 is the one used in the parameter estimation above (see examples 5.3 and 5.4). This subset has a collinearity index of 45, which is far higher than typically considered threshold values of 5-15 for a subset to be considered practically identifiable

(Brun *et al.*, 2002; Sin *et al.,* 2010). As shown here, the analysis would have diagnosed the issue before performing the parameter estimation (PE) and this would have indicated that this subset was not suitable for the estimation.

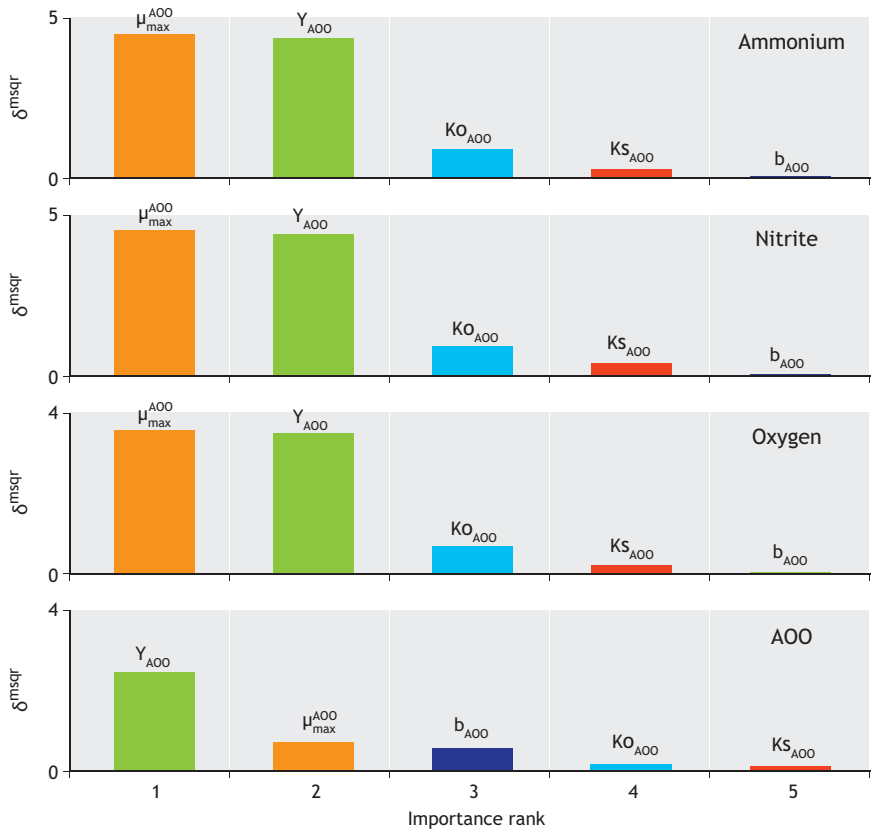

Figure 5.2 Significance ranking of the AOO parameters with respect to the model outputs.

However, given that the sensitivity of b_{AOO} was not influential on the outputs (see Step 3), any subset containing this parameter would not be recommended for parameter estimation. Nevertheless there remain many subsets that meet a threshold of 5-15 for γ_K that can be considered for the parameter estimation problem. The parameter subsets shaded in Table 5.8 meet these identifiability criteria, and therefore can be used for parameter estimation. The best practice is to start with the parameter subset with the largest size (of parameters) and lowest γ_K. Taking these considerations of the sensitivity and collinearity index of the parameter subsets into account helps to avoid the ill-conditioned parameter estimation problem and to improve the quality of the parameter estimates.

Example 5.6 Estimate the model prediction uncertainty of the nitrification model – the Monte Carlo method

In this example, we wish to propagate the parameter uncertainties resulting from parameter estimation (e.g. Example 5.3 and Example 5.4) to model output uncertainty using the Monte Carlo method.

For the uncertainty analysis, the problem is defined as follows: (*i*) only the uncertainty in the estimated AOO parameters is considered, (*ii*) the experimental conditions of batch test 1 are taken in account (Table 5.3), and (*iii*) the model in Table 5.1 is used to describe the system and nominal parameter values in Table 5.2.

Table 5.8 The collinearity index calculation for all the parameter combinations.

Subset K	Subset size	Parameter combination					γ_K
1	2	$K_{o,AOO}$	b_{AOO}				1.32
2	2	$K_{s,AOO}$	b_{AOO}				1.26
3	2	$K_{s,AOO}$	$K_{o,AOO}$				2.09
4	2	μ_{max}^{AOO}	b_{AOO}				1.30
5	2	μ_{max}^{AOO}	$K_{o,AOO}$				13.92
6	2	μ_{max}^{AOO}	$K_{s,AOO}$				2.03
7	2	Y_{AOO}	b_{AOO}				1.28
8	2	Y_{AOO}	$K_{o,AOO}$				12.55
9	2	Y_{AOO}	$K_{s,AOO}$				2.02
10	2	Y_{AOO}	μ_{max}^{AOO}				42.93
11	3	$K_{s,AOO}$	$K_{o,AOO}$	b_{AOO}			2.10
12	3	μ_{max}^{AOO}	$K_{o,AOO}$	b_{AOO}			14.05
13	3	μ_{max}^{AOO}	$K_{s,AOO}$	b_{AOO}			2.03
14	3	μ_{max}^{AOO}	$K_{s,AOO}$	$K_{o,AOO}$			14.23
15	3	Y_{AOO}	$K_{o,AOO}$	b_{AOO}			13.09
16	3	Y_{AOO}	$K_{s,AOO}$	b_{AOO}			2.02
17	3	Y_{AOO}	$K_{s,AOO}$	$K_{o,AOO}$			12.89
18	3	Y_{AOO}	μ_{max}^{AOO}	b_{AOO}			51.25
19	3	Y_{AOO}	μ_{max}^{AOO}	$K_{o,AOO}$			45.87
20	3	Y_{AOO}	μ_{max}^{AOO}	$K_{s,AOO}$			43.37
21	4	Y_{AOO}	μ_{max}^{AOO}	$K_{s,AOO}$	$K_{o,AOO}$		45.91
22	4	Y_{AOO}	μ_{max}^{AOO}	$K_{s,AOO}$	b_{AOO}		51.25
23	4	Y_{AOO}	μ_{max}^{AOO}	$K_{o,AOO}$	b_{AOO}		53.01
24	4	Y_{AOO}	$K_{s,AOO}$	$K_{o,AOO}$	b_{AOO}		13.30
25	4	μ_{max}^{AOO}	$K_{s,AOO}$	$K_{o,AOO}$	b_{AOO}		14.30
26	5	Y_{AOO}	μ_{max}^{AOO}	$K_{s,AOO}$	$K_{o,AOO}$	b_{AOO}	53.07

Step 1. Input uncertainty definition.

As defined in the above problem definition, only the uncertainties in the estimated AOO parameters are taken into account:

- $\theta_{input} = [Y_{AOO}\ \mu_{max}^{AOO}\ K_{s,AOO}\ K_{o,AOO}]$.

Mean and standard deviation estimates are taken as obtained from the bootstrap method together with their correlation matrix (Table 5.7). Further it is assumed that these parameters follow a normal distribution or multivariate normal distribution since they have a covariance matrix and are correlated. This assumption can be verified by calculating the empirical density function for each parameter using the parameter estimates matrix (θ_{50x4}) and shown in Figure 5.12.

```
figure
labels=['\theta_1';'\theta_2';'\theta_3';'\theta
_4']; %or better the name of parameter
for i=1:4
    subplot(2,2,i)
    [f xi]=ksdensity(pmin(:,i));
    plot(xi,f)
xlabel(labels(i,:),'FontSize',fs,'FontWeight','b
old')
end
```

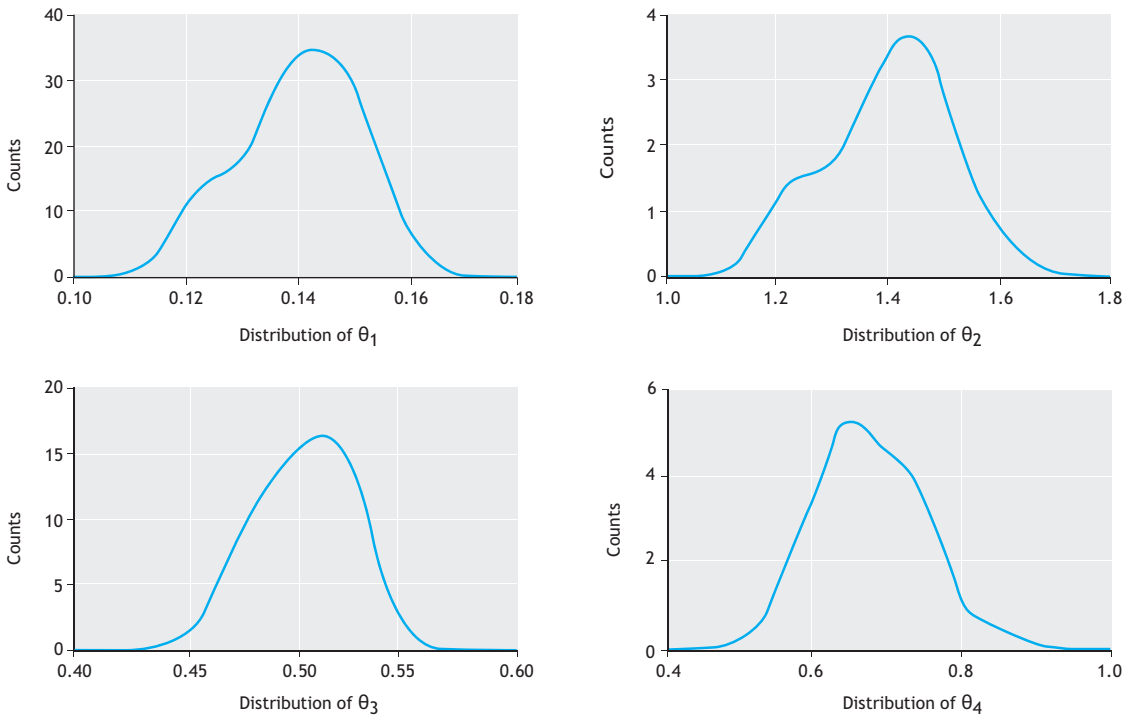

Figure 5.12 Empirical probability density estimates for the AOO parameters as obtained by the bootstrap method.

Step 2. Sampling from the input space

Since the input parameters have a known covariance matrix, any sampling technique must take this into account. In this example, since the parameters are defined to follow a normal distribution, the input uncertainty space is represented by a multivariate normal distribution. A random sampling technique is used to sample from this space:

```
%% do random sampling
N= 100; %% sampling number
mu=mean(pmin); %% mean values of parameters
sigma=cov(pmin); %% covariance matrix (includes
stand dev and correlation information)
X = mvnrnd(mu,sigma,N); % sample parameter space
using multivariate random sampling
```

The output from this step is a sampling matrix, X_{Nxm}, where N is the sampling number and m is the number of inputs. The sampled values can be viewed using a matrix plot as in Figure 5.13. In this figure, which is a matrix plot, the diagonal subplots are the histogram of the parameter values while the non-diagonal subplots show the sampled values of the two pairs of parameters. In this case the most important observations are that (*i*) the parameter input space is sampled randomly and (*ii*) the parameter correlation structure is preserved in the sampled values.

Step 3. Perform the Monte Carlo simulations.

In this step, N model simulations are performed using the sampling matrix from Step 2 (X_{Nxm}) and the model outputs are recorded in a matrix form to be processed in the next step.

```
%%step 2 perform monte carlo simulations for
each parameter value
% Solution of the model
initcond;options=odeset('RelTol',1e-
7,'AbsTol',1e-8);
for i=1:nboot
    disp(['the iteration number is :
',num2str(i)])
    par(idx) = X(i,:) ; %read a sample from
sampling matrix
    [t,y1] = ode45(@nitmod,td,x0,options,par); ;
%solve the model
     y(:,:,i)=y1; %record the outputs
end
```

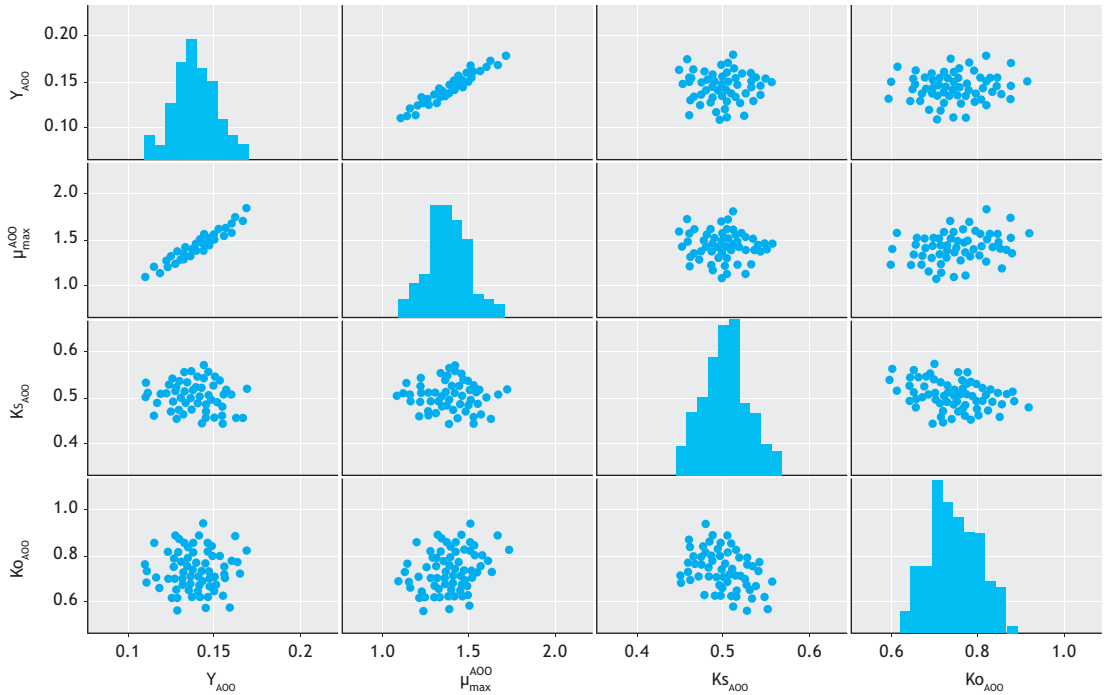

Figure 5.13 Plotting of the sampling matrix of the input space, X_{Nxm} – the multivariate random sampling technique with a known covariance matrix.

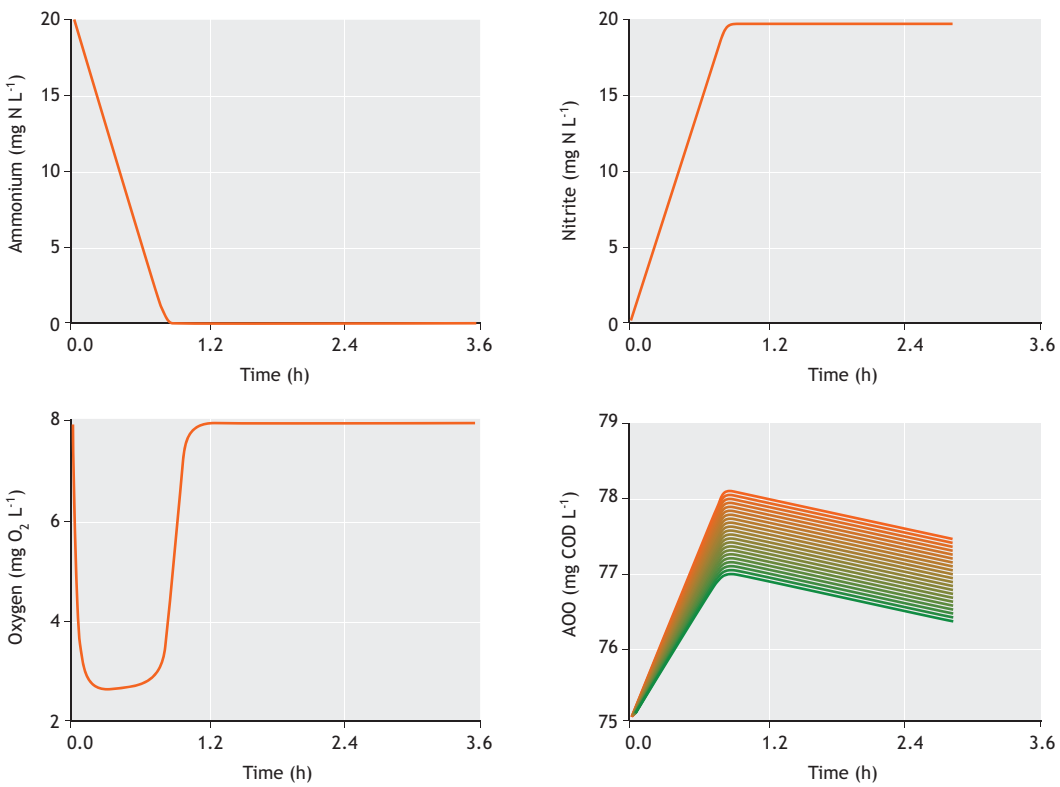

Figure 5.14 Monte Carlo simulations (N = 100) of the model outputs.

Step 4. Review and analyse the results.

In this step, the outputs are plotted and the results are reviewed. In Figure 5.14, Monte Carlo simulation results are plotted for four model outputs.

As shown in Figure 5.15, the mean, standard deviation and percentiles (e.g. 95 %) can be calculated from the output matrix. The results indicate that for the sources of uncertainties being studied, the uncertainty in the model outputs can be considered negligible. These results are in agreement with the linear error propagation results shown in Figure 5.5.

This means that while there is uncertainty in the parameter estimates themselves, when the estimated parameter subset is used together with its covariance matrix, the uncertainty in the model prediction is low. For any application of these model parameters they should be used together as a set, rather than individually.

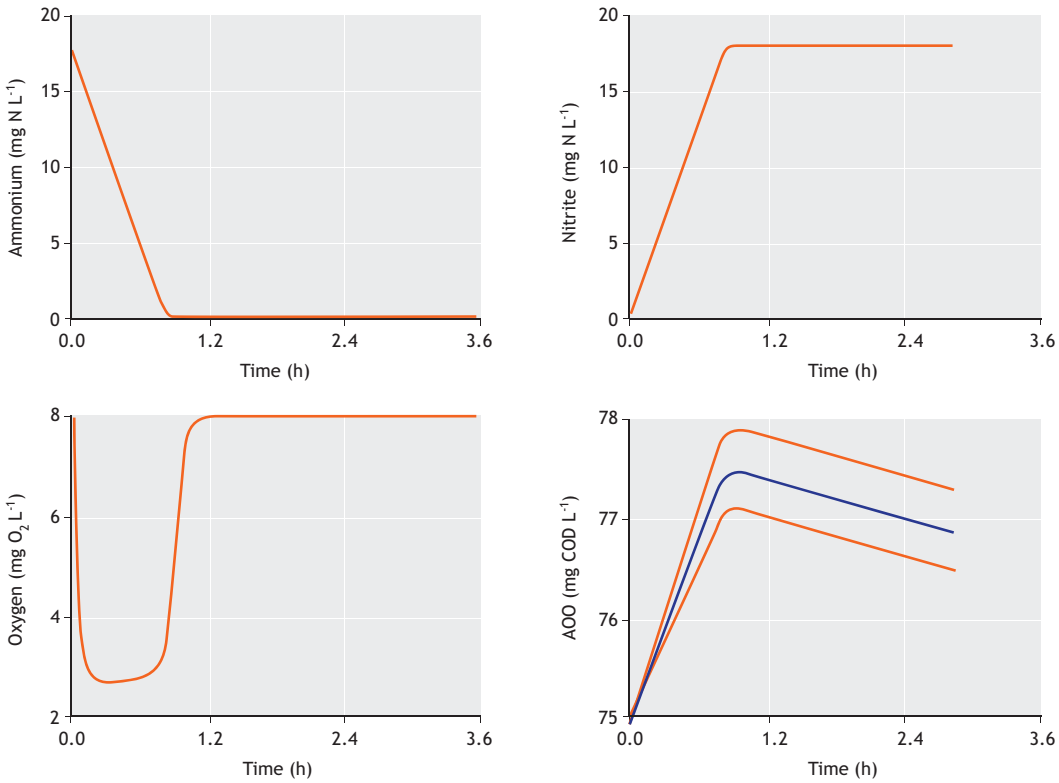

Figure 5.15 Mean and 95 % percentile calculation of the model output uncertainty.

Another point to make is that the output uncertainty evaluated depends on the input uncertainty defined as well as the framing, e.g. initial conditions of the experimental setup. For example, in the above example what was not considered is the measurement uncertainty or uncertainty due to other fixed parameters (decay) and initial conditions (the initial concentration of autotrophic bacteria). Therefore these results need to be interpreted within the context where they are generated.

5.5 ADDITIONAL CONSIDERATIONS

Best practice in parameter estimation

In practice, while asymptotic theory assumption gives reasonable results, there are often deviations from the assumptions. In particular:

- The measurement errors are often auto-correlated, meaning that too many observations are redundant and not independent (non-independently and identically distributed (iid) random variables). This tends to cause an underestimation of asymptotic confidence intervals due to smaller sample variance, σ^2. A practical solution to this problem is to check the autocorrelation function of the residuals and filter them or perform subsampling such that autocorrelation is decreased in the data set. The parameter estimation can then be redone using the subsample data set.
- Parameter estimation algorithms may stop at local minima, resulting in an incorrect linearization result (the point at which the non-linear least squares are linearized). To alleviate this issue, parameter estimation needs to be performed several times with either different initial guesses, different search algorithms and/or an identifiability analysis.

Afterwards it is important to verify that the minimum solution is consistent with different minimization algorithms.

Identifiability or ill-conditioning problem: Not all the parameters can be estimated accurately. This can be caused by a too large confidence interval compared to the mean or optimized value of the parameter estimators. The solution is to perform an identifiability analysis or re-parameterisation of the model, so that a lower number of parameters needs to be estimated.

While we have robust and extensive statistical theories and methods relevant for estimation of model parameters as demonstrated above, the definition of the parameter estimation problem itself, which is concerned with stating what is the data available, what is the candidate model structure, and what is the starting point for the parameter values, is taken for granted. Hence a proper analysis and definition of the parameter estimation problem will always require a good engineering judgment. For robust parameter estimation in practice, due to the empirical/experiental nature of parameter definition, the statistical methods (including MLE estimates) should be treated within the context/definition of the problem of interest.

With regards to bootstrap sampling, the most important issue is whether or not the residuals are representative of typical measurement error. For a more detailed discussion of this issue, refer to Efron (1979).

Best practice in uncertainty analysis

When performing uncertainty analysis, the most important issue is the framing and the corresponding definition of the input uncertainty sources. Hence, the outcome from an uncertainty analysis should not be treated as absolute but dependent on the framing of the analysis. A detailed discussion of these issues can be found elsewhere (e.g. Sin *et al.,* 2009; Sin *et al.,* 2010).

Another important issue is the covariance matrix of the parameters (or correlation matrix), which should be obtained from a parameter estimation technique. Assuming the correlation matrix is negligible may lead to over or under estimation of the model output uncertainty. Hence, in a sampling step the appropriate correlation matrix should be defined for inputs (*e.g.* parameters) considered for the analysis.

Regarding the sampling number, one needs to iterate several times to see if the results differ from one iteration to another. Since the models used for the parameter estimation are relatively simple to solve numerically, it is recommended to use a sufficiently high number of iterations e.g. 250 or 500.

References

Brun, R., Kühni, M., Siegrist, H., Gujer, W., Reichert, P. (2002). Practical identifiability of ASM2d parameters - systematic selection and tuning of parameter subsets. *Water Res.* 36(16): 4113-4127.

Brun, R., Reichert, P., and Künsch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 37(4):1015-1030.

Bozkurt, H., Quaglia, A., Gernaey, K.V., Sin, G. (2015). A mathematical programming framework for early stage design of wastewater treatment plants. *Environmental Modelling & Software*, 64: 164-176.

Dochain, D., Vanrolleghem, P.A. (2001). Dynamical Modelling and Estimation in Wastewater Treatment Processes. London UK: IWA Publishing.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1-26.

Gernaey, K.V., Jeppsson, U., Vanrolleghem, P.A., Copp, J.B. (Eds.). (2014). *Benchmarking of control strategies for wastewater treatment plants*. IWA Publishing.

Guisasola, A., Jubany, I., Baeza, J.A., Carrera, J., Lafuente, J. (2005). Respirometric estimation of the oxygen affinity constants for biological ammonium and nitrite oxidation. *Journal of Chemical Technology and Biotechnology*, 80(4): 388-396.

Heijnen, J.J. (1999). Bioenergetics of microbial growth. *Encyclopaedia of Bioprocess Technology*.

Henze, M., Gujer, W., Mino, T., van Loosdrecht, M.C.M., (2000). ASM2, ASM2d and ASM3. *IWA Scientific and Technical Report*, *9*. London UK.

Ljung L. (1999). System identification - Theory for the user. 2nd edition. Prentice-Hall.

Mauricio-Iglesias, M., Vangsgaard, A.K., Gernaey, K.V., Smets, B.F., Sin, G. (2015). A novel control strategy for single-stage autotrophic nitrogen removal in SBR. *Chemical Engineering Journal*, 260: 64-73.

Meijer, S.C.F., Van Der Spoel, H., Susanti, S., Heijnen, J.J., van Loosdrecht, M.C.M. (2002). Error diagnostics and data reconciliation for activated sludge modelling using mass balances. *Water Sci Tech.* 45(6): 145-156.

Metropolis, N., Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247): 335-341.

Omlin, M. and Reichert, P. (1999). A comparison of techniques for the estimation of model prediction uncertainty. *Ecol. Model.,* 115: 45-59.

Roels, J.A. (1980). Application of macroscopic principles to microbial metabolism. *Biotechnology and Bioengineering*, 22(12): 2457-2514.

Saltelli, A., Tarantola, S., and Campolongo, F. (2000). Sensitivity analysis as an ingredient of modeling. *Statistical Science,* 15(4):377-395.

Seber G. and Wild C. (1989) Non-linear regression. Wiley, New York.

Sin, G., Gernaey, K.V., Neumann, M.B., van Loosdrecht, M.C.M., Gujer, W. (2009). Uncertainty analysis in WWTP model applications: a critical discussion using an example from design. *Water Res.* 43(11): 2894-2906.

Sin, G., Gernaey, K.V., Neumann, M.B., van Loosdrecht, M.C.M., Gujer, W. (2011). Global sensitivity analysis in wastewater treatment plant model applications: prioritizing sources of uncertainty. *Water Research,* 45(2): 639-651.

Sin, G., de Pauw, D.J.W., Weijers, S., Vanrolleghem, P.A. (2008). An efficient approach to automate the manual trial and error calibration of activated sludge models. *Biotechnology and Bioengineering.* 100(3): 516-528.

Sin, G., Meyer, A.S., Gernaey, K.V. (2010). Assessing reliability of cellulose hydrolysis models to support biofuel process design-identifiability and uncertainty analysis. *Computers & Chemical Engineering*, 34(9): 1385-1392.

Sin, G., Vanrolleghem, P.A. (2007). Extensions to modeling aerobic carbon degradation using combined respirometric–titrimetric measurements in view of activated sludge model calibration. *Water Res.* 41(15): 3345-3358.

Vangsgaard, A.K., Mauricio-Iglesias, M., Gernaey, K.V., Sin, G. (2014). Development of novel control strategies for single-stage autotrophic nitrogen removal: A process oriented approach. *Computers & Chemical Engineering*, 66: 71-81.

van der Heijden, R.T.J.M., Romein, B., Heijnen, J.J., Hellinga, C., Luyben, K. (1994). Linear constraint relations in biochemical reaction systems: II. Diagnosis and estimation of gross errors. *Biotechnology and bioengineering*, 43(1): 11-20.

Villadsen, J., Nielsen, J., Lidén, G. (2011). Elemental and Redox Balances. In *Bioreaction Engineering Principles* (pp. 63-118). Springer, US.