Sequential Parameter Estimation for Mammalian Cell Model Based on In Silico Design of Experiments

Authors:

Zhenyu Wang, Hana Sheikh, Kyongbum Lee, Christos Georgakis

Date Submitted: 2018-08-28

Keywords: Design of Experiments, sensitivity analysis, parameter estimation, Mammalian Cell Culture, Pharmaceutical Processes

Abstract:

Due to the complicated metabolism of mammalian cells, the corresponding dynamic mathematical models usually consist of large sets of differential and algebraic equations with a large number of parameters to be estimated. On the other hand, the measured data for estimating the model parameters are limited. Consequently, the parameter estimates may converge to a local minimum far from the optimal ones, especially when the initial guesses of the parameter values are poor. The methodology presented in this paper provides a systematic way for estimating parameters sequentially that generates better initial guesses for parameter estimation and improves the accuracy of the obtained metabolic model. The model parameters are first classified into four subsets of decreasing importance, based on the sensitivity of the model's predictions on the parameters' assumed values. The parameters in the most sensitive subset, typically a small fraction of the total, are estimated first. When estimating the remaining parameters with next most sensitive subset, the subsets of parameters with higher sensitivities are estimated again using their previously obtained optimal values as the initial guesses. The power of this sequential estimation approach is illustrated through a case study on the estimation of parameters in a dynamic model of CHO cell metabolism in fed-batch culture. We show that the sequential parameter estimation approach improves model accuracy and that using limited data to estimate low-sensitivity parameters can worsen model performance.

Record Type: Published Article

Submitted To: LAPSE (Living Archive for Process Systems Engineering)

Citation (overall record, always the latest version):	LAPSE:2018.0404
Citation (this specific file, latest version):	LAPSE:2018.0404-1
Citation (this specific file, this version):	LAPSE:2018.0404-1v1

DOI of Published Version: https://doi.org/10.3390/pr6080100

processes

MDPI

Article

Sequential Parameter Estimation for Mammalian Cell Model Based on In Silico Design of Experiments

Zhenyu Wang, Hana Sheikh, Kyongbum Lee and Christos Georgakis *

Department of Chemical and Biological Engineering and Systems Research Institute for Chemical and Biological Processes Tufts University, Medford, MA 02155, USA; zwang12@dow.com (Z.W.); Hana.Sheikh@gmail.com (H.S.); Kyongbum.Lee@tufts.edu (K.L.)

* Correspondence: christos.georgakis@tufts.edu; Tel.: +1-617-627-2573

Received: 18 May 2018; Accepted: 20 July 2018; Published: 24 July 2018

Abstract: Due to the complicated metabolism of mammalian cells, the corresponding dynamic mathematical models usually consist of large sets of differential and algebraic equations with a large number of parameters to be estimated. On the other hand, the measured data for estimating the model parameters are limited. Consequently, the parameter estimates may converge to a local minimum far from the optimal ones, especially when the initial guesses of the parameter values are poor. The methodology presented in this paper provides a systematic way for estimating parameters sequentially that generates better initial guesses for parameter estimation and improves the accuracy of the obtained metabolic model. The model parameters are first classified into four subsets of decreasing importance, based on the sensitivity of the model's predictions on the parameters' assumed values. The parameters in the most sensitive subset, typically a small fraction of the total, are estimated first. When estimating the remaining parameters with next most sensitive subset, the subsets of parameters with higher sensitivities are estimated again using their previously obtained optimal values as the initial guesses. The power of this sequential estimation approach is illustrated through a case study on the estimation of parameters in a dynamic model of CHO cell metabolism in fed-batch culture. We show that the sequential parameter estimation approach improves model accuracy and that using limited data to estimate low-sensitivity parameters can worsen model performance.

Keywords: Pharmaceutical Processes; Mammalian Cell Culture; sensitivity analysis; parameter estimation; Design of Experiments

1. Introduction

The use of biologics, including antibiotics and antibodies, has increased across different therapeutic areas, and is poised to fuel pharmaceutical revenues and stimulate growth in the biopharmaceutical market. In 2012, the global sales of biologics reached 124.9 billion in US dollars, a 10.4% increase over 2011 [1]. More than half of the therapeutic recombinant proteins are produced in immortalized mammalian cell lines, including Chinese hamster ovary (CHO), baby hamster kidney (BHK), and mouse myeloma cells (NS0). Dynamic models of cellular metabolism have been developed to provide insight into the mechanism behind a process, and further enable prediction and optimization for the productivity of cell cultures [2–5]. These metabolic models in some cases contain hundreds of rate constants to describe the rate processes occurring in the cell and bioreactor. Very often, the experimental data set is not large enough to allow for the estimation of all the parameters in the metabolic model [6]. Usually, only a subset of the parameters might be estimable [7]. Moreover, parameter estimation presents a difficult challenge due to the complicated, nonlinear structure of the metabolic models. This problem is exacerbated when some of the parameters have little impact on the

model's outputs. A method to systematically select the subset of parameters with the largest impact on the model outputs would greatly benefit not only parameter estimation, but also model validation.

Sensitivity analysis methodologies, including local and global sensitivity analysis, are widely applied to select the subset of parameters with the largest impact on the outputs of a metabolic model to be estimated with available data [8]. The local sensitivity analysis (LSA) approaches calculate the sensitivity coefficients via partial derivatives of output variables with respect to each model parameter with given values of input variables and nominal values of other parameters. Once the sensitivity matrix with sensitivity coefficients as elements has been constructed, a variety of methods, including orthogonalization algorithm [9,10], the Mean Squared Error-based method [11,12], and the Principal Component Analysis-based method [13], can be applied to rank the importance of the parameters. The recent progresses in LSA has been well-reviewed by [7]. As the sensitivity coefficients are calculated based on the impacts by individual parameters, the LSA approaches do not account for the interaction impact of multiple parameters, which may significantly affect the output variables of nonlinear and complicated models.

On the other hand, Global Sensitivity Analysis (GSA), also known as Sobol's method [14], calculates the sensitivity coefficients by simultaneously varying all the parameters in the range of interest. The obtained sensitivity coefficients account for the interaction impact of all the parameters, as well as the impact of individual parameters. GSA is frequently employed to identify the most important and sensitive parameters of metabolic models [15–17]. The major limitation of such a method based on Monte Carlo simulations is that it incurs high computational cost. To reduce the computational effort, [18] proposed using a meta-model, a Response Surface Methodology (RSM) model [19], to approximate the original dynamic model comprising a large set of differential and algebraic equations. Then, the sensitivity analysis is conducted on the simplified meta-model via Monte Carlo simulations similar to GSA. As the meta-model is an algebraic model, the computational cost for simulating such a model is drastically less than a dynamic model of cellular metabolism. Consequently, the sensitivity analysis is completed much faster.

In this paper, we present a new sequential parameter estimation methodology, prioritizing the estimation of parameters with descending sensitivity indices. We first improved the sensitivity analysis approach proposed by [18] by eliminating the step of running Monte Carlo simulations on the meta-model. Instead, we calculate the sensitivity index analytically using the estimated RSM model. Moreover, we propose a systematic approach to discriminate the parameters into four categories based on the sensitivity of the model outputs. This allows the modeler to prioritize the estimation of the model parameters by initially focusing on those with the highest importance or largest sensitivity indices. We demonstrate the power of the proposed method by successfully identifying the important parameters in a well-received model of CHO cell metabolism [4] using experimental data. We show that our sequential parameter estimation method results in a more accurate model compared to when all of the parameters are estimated simultaneously.

2. Global Sensitivity Analysis

Sobol's method, or Global Sensitivity Analysis (GSA), varies the parameters of interest simultaneously over their entire domain to examine the interaction effects among parameters. The dynamic model of interest is decomposed into sums of orthogonal functions (also known as summand), as given below.

$$y = f(\boldsymbol{\theta}) = g_0 + \sum_{s=1}^{n} \sum_{i_1 < i_2 < \cdots < i_s}^{n} g_{i_1 \cdots i_s}(\theta_{i_1}, \cdots, \theta_{i_s}) \tag{1}$$

where y is the output variable, while $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_n \end{bmatrix}^T$ is a column vector with n parameters as elements. Each orthogonal function $g(\cdot)$ represents the effect of corresponding

parameters on the output. If the output, y, is affected by two parameters, θ_1 and θ_2, the expansion of Equation (1) is given by

$$y = g_0 + g_1(\theta_1) + g_2(\theta_2) + g_{12}(\theta_1, \theta_2) \tag{2}$$

The magnitude of the effect on the output by each variable is related to the variance of the corresponding orthogonal function, $g(\cdot)$, which is calculated as follows:

$$\begin{aligned} D &= \int_{\boldsymbol{a}}^{\boldsymbol{b}} f(\boldsymbol{\theta})^2 d\boldsymbol{\theta} - g_0^2 \\ D_i &= \int_{a_i}^{b_i} g_i^2(\theta_i) d\theta_i \\ D_{i_1 \cdots i_s} &= \int_a^b g_{i_1 \cdots i_s}^2 \left(\theta_{i_1 \cdots i_s}\right) d\theta_{i_1 \cdots i_s} \end{aligned} \tag{3}$$

where $D = \sum_{s=1}^{n} \sum_{i_1 < \cdots < i_s}^{n} D_{i_1 \cdots i_s}$ represents the total variance in outputs due to all parameters in the domain defined by $[\boldsymbol{a}, \boldsymbol{b}]$. The $n \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are the lower and upper bounds for the n parameters, while D_i represents the variance in outputs due to parameter θ_i in the corresponding range of $[a_i,\ b_i]$. The sensitivity of an output to a parameter is quantified by the Total Sensitivity Index (TSI), as given below

$$TSI_i = \frac{D_i + \sum_{j=1}^{n} D_{ij} + \sum_{j=1}^{n} \sum_{k=1}^{n} D_{ijk} + \cdots}{D} \tag{4}$$

The larger the TSI_i, the more strongly the corresponding parameter affects the output. Therefore, a parameter is considered more important if its TSI is larger. For a complicated model comprising a system of many nonlinear differential equations, the explicit solution for the summands, $g(\cdot)$, cannot be obtained. The corresponding variance, and therefore the sensitivity indices, are instead estimated through Monte Carlo simulations, e.g., using Satelli's algorithm [20].

3. Sensitivity Analysis Based on In Silico Design of Experiments

To reduce the computational cost, the method proposed by [18] first estimates a Response Surface Methodology (RSM) model, and then determines the sensitivity index by running Monte Carlo Simulations on the estimated RSM model. To estimate the RSM model, the parameters, $\boldsymbol{\theta}$, are first coded in the range of $[-1,\ +1]$. For each model parameter θ_i, the corresponding coded parameter x_i is given by Equation (5).

$$x_i = (\theta_i - \theta_{0,i}) / \Delta\theta_i \tag{5}$$

Here, $\theta_{0,i}$ is the reference value of parameter θ_i and $\Delta\theta_i$ is the half interval in which we expect the parameter's optimal value will lie. Both θ_i and $\Delta\theta_i$ are selected by the modeler based on the modeler's understanding of the process. If the estimated value of the parameter is at an endpoint of this interval, the initial choice of the interval might need to be corrected.

To minimize the number of detailed simulations, a D-optimal design [19] of in silico experiments is performed, where the model parameters are systematically varied in the range of $\theta_0 \pm \Delta\theta$ around their nominal values θ_0. Customarily, these inputs are transformed into their dimensionless coded form as defined above. The resulting time-resolved output values, $y(t|\boldsymbol{x})$, for the defined combinations of the coded inputs, $\boldsymbol{x}$, are collected through the simulation of the metabolic model and are used to estimate the RSM model of $y(t|\boldsymbol{x})$ to $\boldsymbol{x}$. An example of a quadratic RSM model with n coded inputs or factors is given by

$$y(t|\boldsymbol{x}) = f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{j>i}^{n} \beta_{ij} x_i x_j + \sum_{i=1}^{n} \beta_{ii} x_i^2 \tag{6}$$

The β_i, β_{ij}, and β_{ii} are coefficients of the RSM model and are estimated by stepwise regression [21] to avoid overfitting of the model. Instead of conducting a GSA on the complicated system of differential equations, the GSA is applied to the relatively simple algebraic RSM model as given above for each of the outputs of interest. Specifically, sensitivity indices are calculated using the outputs obtained by

varying the coded inputs simultaneously. As shown in [18], the Design of Experiment-based method significantly reduces the required computational time.

4. Sequential Parameter Estimation

Here we present a new sequential parameter estimation method consisting of an improved method for the aforementioned sensitivity analysis approach and a systematic way to prioritize the estimation of subsets of parameters. We first improve the above sensitivity analysis method by eliminating the step of running Monte Carlo simulations on the estimated RSM model. In the improved method, the sensitivity index of each parameter is estimated analytically following Sobol's method, as detailed below. We here illustrate the proposed approach using a quadratic RSM model, given in Equation (6), as an example. If a higher-order RSM model is at hand, the sensitivity index of each parameter will be determined in a similar manner.

Using the definition given by Sobol [14], the orthogonal summands can be solved as follows:

$$\begin{aligned} g_0 &= \tfrac{1}{L^n}\int_{-1}^{+1} f(\boldsymbol{x})d\boldsymbol{x} = \beta_0 + \tfrac{1}{3}\sum_{i=1}^{n}\beta_{ii} \\ g_i(x_i) &= \tfrac{1}{L^{n-1}}\int_{-1}^{+1} f(\boldsymbol{x})dx_{k\neq i} - g_0 = \beta_i x_i + \beta_{ii}(x_i^2 - \tfrac{1}{3}) \\ g_{i,j}(x_i, x_j) &= \tfrac{1}{L^{n-2}}\int_{-1}^{+1} f(\boldsymbol{x})dx_{k\neq i,j} - g_0 - g_i - g_j = \beta_{ij}x_i x_j \end{aligned} \tag{7}$$

Here L is the range of the input variables in the RSM model. The input variables here are the parameters of the metabolic model. These are coded into the range of $[-1, +1]$. We use $L = 2$ to derive the orthogonal functions as shown in Equation (7). By substituting the orthogonal functions into Equation (3), we express the variance functions as follows:

$$\begin{aligned} D_i &= \tfrac{1}{L}\int_{-1}^{+1} g_i^2(x_i)dx_i = \tfrac{1}{3}\beta_i^2 + \tfrac{4}{45}\beta_{ii}^2 \\ D_{i,j} &= \tfrac{1}{L^2}\int_{-1}^{+1} g_{i,j}^2(x_i,\ x_j)dx_i dx_j = \tfrac{1}{9}\beta_{ij}^2 \end{aligned} \tag{8}$$

Then the total variance is calculated as $D = \sum_{s=1}^{n}\sum_{i_1<\cdots<i_s}^{n} D_{i_1\cdots i_s}$. The total sensitivity indices are calculated by substituting the variances calculated above into Equation (4). As the quadratic RSM model accounts for the interaction effect of up to two inputs, the sensitivity index is calculated as follows.

$$TSI_i = \frac{D_i + \sum_{j=1}^{n} D_{ij}}{D} \tag{9}$$

The results in Equations (8) and (9) can be easily generalized to estimate higher order (>2) sensitivity indices. However, this will require the estimation of RSM of higher order.

Each parameter is ranked from most to least important, according to its sensitivity index. Next, we classify the parameters into three subsets, most important (subset A), important (subset B) and least important (subset C), based on the percentage of explained output variance. We would like subset A to explain at least α% of the total output variance, subset B to explain an additional β%, and subset C to explain another γ% of the variance. The restriction is that $(\alpha + \beta + \gamma) < 100\%$. This could leave out a small percentage of the overall variance, i.e., $100\% - (\alpha + \beta + \gamma)$. This small percentage typically corresponds to the noise in the data and, thus, is of low importance in terms of parameter estimation. The set of parameters that are not selected in the subsets A, B and C are grouped into subset D. In the present study, we set the values of α, β, and γ to 50%, 30% and 10%, respectively. The remaining unexplained variance, $100\% - (\alpha + \beta + \gamma)$, is then 10% of the overall and is related to the parameters in subset D, i.e., parameters which have the smallest effect on the model's predictions and whose values it might not be worth adjusting further beyond their initial estimates. More generally, the specific values of the parameters α, β, and γ are left to the discretion of the modeler, as well as the $100\% - (\alpha + \beta + \gamma)$

fraction of the variance that could have been addressed by minute adjustments to a possibly large number of parameters, each one of which has a negligible impact on the model predictions.

Sobol's sensitivity index represents the ratio of the output variances caused by a certain variable change to the output variances caused by all variable changes. Therefore, we can quantify the output variance by a given subset of parameters by directly summing the corresponding sensitivity indices. Given a set of N total parameters and a subset of S parameters of interest, the percentage of variance is calculated using the following equation.

$$\eta = \frac{\sum_{i \in S} TSI_i}{\sum_{j=1}^{N} TSI_j} \times 100\% \tag{10}$$

Using Equation (10) and the selected values for the thresholds α, β, and γ, we divide the parameters into four subsets, A, B, C, and D. We then sequentially estimate the values of the parameter's starting with those in subset A, then those in subsets A and B and finally those in subsets A, B and C. When we estimate the most important subset of parameters, subset A, we hold the remaining parameters in subset B, C and D at their nominal values. These nominal values can be obtained from the literature or approximately estimated based on available knowledge about the process. They will most likely be equal to the reference values, θ_0, defined above. This reduces the dimensionality of the optimization problem that has to be solved in each parameter estimation task, substantially alleviating the challenge caused by local minima, especially prevalent when the number of decision variables is very large. The parameter estimation problem can be defined mathematically as follows:

$$\theta_S^* = \underset{\theta_S}{\operatorname{argmin}} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\frac{\hat{y}_{m,k}(\theta_S, \theta_{i \notin S} = \theta_0) - y_{m,k}}{y_{m,k}} \right)^2 \tag{11}$$

where θ_S are the parameters to be estimated and $\theta_{i \notin S}$ are the parameters to be held at their fixed nominal values, θ_0. Also $\hat{y}_{m,\,k}$ and $y_{m,k}$ are the values predicted by the model and the corresponding measured values of species m at time instant k. The above parameter estimation problem is solved in Matlab's Optimization toolbox [22] using an interior-point algorithm [23] and *fmincon* function. Once the most important parameters in subset A have been estimated, we can estimate the values of parameters in subset B. In this round of estimation, we will take the optimal values for the parameters in subset A, $\theta_{S \in A}^*$, and the nominal values for parameters in subset B, $\theta_{S \in B}$, as the initial guess to solve the optimization problem given in Equation (10). Note that the newly estimated parameters in subset A may be slightly different from the originally estimated nominal values. Next, the parameters in subset C are estimated together with subset A and B parameters using a similar approach. The initial guesses for subset C parameters are their nominal values while the initial guesses for subset A and B parameters are their optimal values obtained in the previous round of parameter estimation. By sequentially estimating the subsets of parameters of descending importance, we obtain a dynamic model of improved accuracy compared to the model where all parameters are estimated simultaneously.

5. Results and Discussion

In this section, we apply the proposed method to estimate the model parameters of a complicated dynamic model of CHO cell metabolism in a fed-batch reactor [4]. The CHO cell model consists of 34 reactions and 51 parameters, and covers major pathways of central carbon metabolism. The model explicitly accounts for redox- and temperature-dependent changes to pathway activities, and directly calculates the measured variables, i.e., metabolite concentration time profiles in the reactor, by defining rate expressions based on extracellular metabolites. The reactions and metabolites involved in the model are visualized in Figure 1.

There are two types of parameters of the CHO cell model: four parameters related to the process operation and 47 kinetic parameters. The process parameters are the shift temperature, shift day, seed density and harvest day. These parameters relate to the operation of the reactor and are selected

based on the choice of the typical operational ranges. In this work, we fix the process parameter ranges to their default values [4] and estimate only the kinetic parameters. However, to identify which process parameters significantly affect the metabolites, we conduct sensitivity analysis for the process parameters as well. The intervals over which the process parameters (shift temperature, shift day, seed density and harvest day) will be varied are 31 ± 3 , 3 ± 1 day, $(3.6 \pm 1.8) \times 10^6$ cell/mL, and 9 ± 1 day, respectively. We label the process parameters #1 to #4.

In Table 1, we define parameters #5 to #51, which refer to the kinetic parameters that have to be estimated from the experimental data. This table also identifies which model parameters are assigned to subsets A, B, C, and D. The 47 kinetic parameters include the maximal reaction velocities (v_{max}), the half saturation constants (K_m), the inhibition constants (K_i) and the temperature dependency constants (TC). The nominal values for the kinetic parameters are set to the values defined in the original model [4]. For the present analysis, these parameters are scaled to be in the same order of magnitude by multiplying each of the parameters with a corresponding scaling factor as given below.

$$\theta_{0,i} = \mathrm{c}\theta_i' \tag{12}$$

where θ_i' is the i^{th} parameter in the original model, while $\theta_{0,i}$ is the corresponding scaled parameter and c is the scaling factor. The values of c and $\theta_{0,i}$ are given in columns 2 and 3 of Table 1, respectively. In each simulation-based sensitivity analysis experiment, the kinetic parameters are varied by $\pm$20% of the nominal values. To estimate the linear and nonlinear sensitivities of the 51 parameters, we design a set of 1398 experiments using the D-Optimal design. Of these, 1378 runs are used to estimate the parameters in a quadratic RSM model, 10 center point runs to represent the expected fed-batch process variability, and 10 additional runs to estimate the Lack-of-Fit (LoF) statistics. A 4% normally distributed error is added to all simulation results to reflect the expected normal variability of the process. This is also the accuracy we expect of the model.

Table 1. Parameter values of CHO Cell model obtained via simultaneous and sequential parameter estimation.

Parameter	Scaling Factor	Scaled Parameter	Subset A	Subset A + B	Subset A + B + C	Simultaneous
V_{max1} (#05)	0.001	3.8		3.75	3.77	3.46
K_{i1} (#06)	0.1	2				1.78
K_{m1} (#07)	0.1	1		1.02	1.01	0.87
Exp_{1a} (#08)	1	3		3.02	3.02	2.67
Exp_{1b} (#09)	1	1				1.11
TB_{1b} (#10)	1	5				4.39
V_{max2} (#11)	0.001	2.2				2.12
K_{m2} (#12)	1	6			6.00	5.73
V_{max3f} (#13)	0.01	3.5	4.10	4.06	4.07	3.97
V_{max3r} (#14)	0.01	1.5			1.51	1.36
K_{m3a} (#15)	1	4		3.97	3.99	3.49
K_{m3b} (#16)	1	2.5				2.16
K_{m3c} (#17)	1	5			5.04	4.55
TC_3 (#18)	1	2				2.18
V_{max8f} (#19)	0.001	2.2		2.24	2.25	2.21
V_{max8r} (#20)	0.01	2	2.31	2.26	2.26	1.97
K_{m8a} (#21)	1	2.5		2.51	2.49	2.27
K_{m8b} (#22)	1	1			0.99	0.95
K_{m8c} (#23)	1	1		1.00	1.01	1.11
TC_{8b} (#24)	1	5				5.33
V_{max9f} (#25)	1	1				1.07
V_{max9r} (#26)	1	1				0.87
K_{m9} (#27)	10	7		7.06	7.06	6.06
V_{max10f} (#28)	0.01	4.75		4.79	4.82	4.45
V_{max10r} (#29)	0.1	2			2.01	2.21
K_{m10z} (#30)	10	3		3.01	3.02	3.06
K_{m10b} (#31)	1	1				0.87
K_{m10c} (#32)	1	2			2.01	1.73
TC_{10b} (#33)	1	1.5		1.49	1.49	1.30
V_{max11} (#34)	10	5.5		5.55	5.58	5.69

Table 1. *Cont.*

Parameter	Scaling Factor	Scaled Parameter	Subset A	Subset A + B	Subset A + B + C	Simultaneous
V_{max12f} (#35)	10	0.9		0.90	0.89	1.00
V_{max12r} (#36)	0.1	2.5				2.35
K_{m12a} (#37)	1	1		0.99	0.99	1.11
K_{m12b} (#38)	1	3				2.60
V_{max13} (#39)	0.1	3				3.33
K_{m13} (#40)	1	1			1.01	1.07
V_{max16} (#41)	0.001	2.5	2.93	2.91	2.91	2.84
K_{m16a} (#42)	10	4		4.00	4.03	4.44
K_{m16b} (#43)	10	3		3.04	3.06	3.20
K_{m16c} (#44)	0.1	2		2.02	2.01	2.22
TC_{16b} (#45)	1	3				3.09
V_{max17} (#46)	0.01	5.25	5.63	5.67	5.69	5.82
K_{i17} (#47)	0.1	3				2.80
Exp_{17a} (#48)	10	5				5.64
Exp_{17b} (#49)	1	1				1.06
V_{max33a} (#50)	10	2				2.22
V_{max33b} (#51)	10	2		2.03	2.01	1.96

The model calculates values for 48 outputs: 34 reaction or exchange fluxes, 14 external metabolite concentrations, including biomass and antibody titer. Since the reaction and exchange fluxes depend on the metabolite concentrations, and the total cell density is directly proportional to biomass, there are only 14 independent outputs. Therefore, 14 quadratic RSMs are developed for the 14 independent metabolite concentrations. The inputs, or factors, in these RSM models are the 51 parameters (4 process and 47 kinetic constants) of the metabolic model whose sensitivity we are trying to assess. The same fractional error is added to the 10 replicated center point runs, through which the Analysis of Variance (ANOVA) [19] estimates the normal variability of the process.

Using the 14 RSMs, the sensitivity indices are calculated by applying Equations (8) and (9). For the case of a single output model, one can simply rank the importance of the parameters according to their sensitivity indices. For the case of multiple outputs, there could be different rankings for each parameter with respect to different outputs. Thus, we need to consider the importance of a parameter to multiple outputs and determine its overall importance. We separate the outputs into two classes: product class and product-relevant class. The product class comprises only the product output itself. Based on the sensitivity of each parameter with respect to the product output, we rank the parameters into four subsets as described previously. The parameters ranked in this fashion are assigned into subsets A1, B1, C1 and D1, in the order of decreasing importance. We then perform the parameter rankings again, this time based on the *average* sensitivities of the parameters with respect to the product-relevant outputs, and assign the parameters into subsets A2, B2, C2 and D2. We combine subsets A1 and A2 to form subset A, which contains the most important parameters overall. Subsets B and C are obtained similarly by combining the corresponding subsets (B1 and B2 and C1 and C2) subject to the condition that the parameters already assigned to a more important subset are excluded. For example, the subset B is the union of subsets B1 and B2 but excludes parameters that are in subset A.

In the CHO cell model case study, the antibody is the desired product. Therefore, we assign this model output to the product class. The remaining model outputs, i.e., metabolites such as glucose and lactate, are either substrates utilized by the cell or highly correlated with the antibody. These model outputs are assigned to the product-relevant class. We calculate the average of sensitivity indices of the kinetic parameters with respect to these outputs and rank the parameters. We apply the aforementioned described threshold criteria to identify the two most important subsets of parameters, A1 and A2 for two classes of outputs that contribute to the desired α fraction (50%) of the variance. The two subsets are combined to define subset A. Subsets B, C, and D are obtained in a similar manner.

The sensitivity indices of the parameters on the antibody concentration and the averaged sensitivity indices on the remaining model outputs are plotted in Figure 2a,b, respectively. In general,

the outputs are strongly sensitive to all four process parameters. In comparison, the sensitivities of the outputs to the kinetic parameters are more varied.

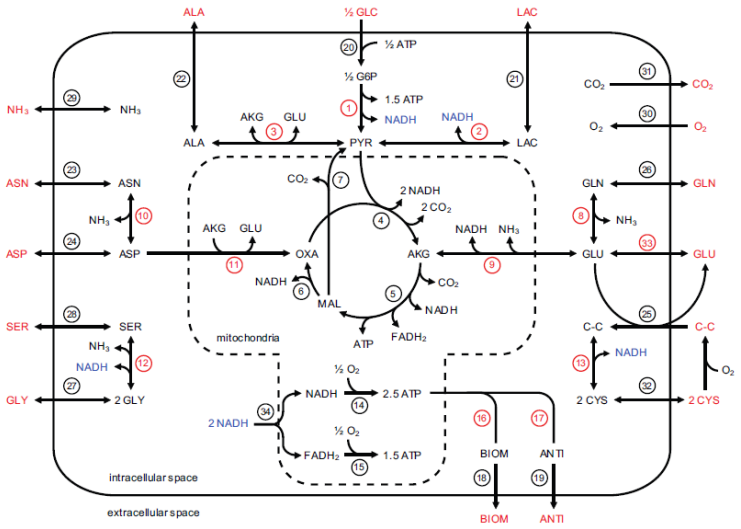


Figure 1. Metabolic pathways of CHO cell model. The 34 reactions are indexed. 14 extracellular metabolites, except for CO_2 and O_2, are modeled.

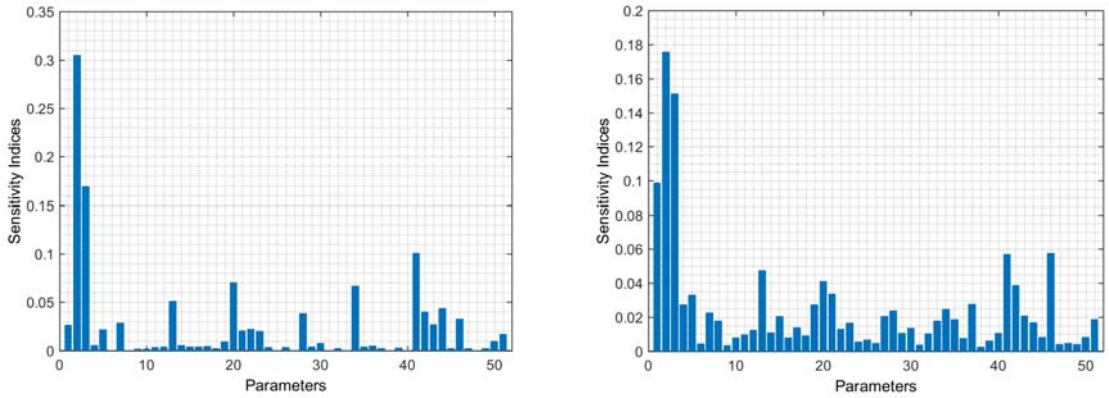
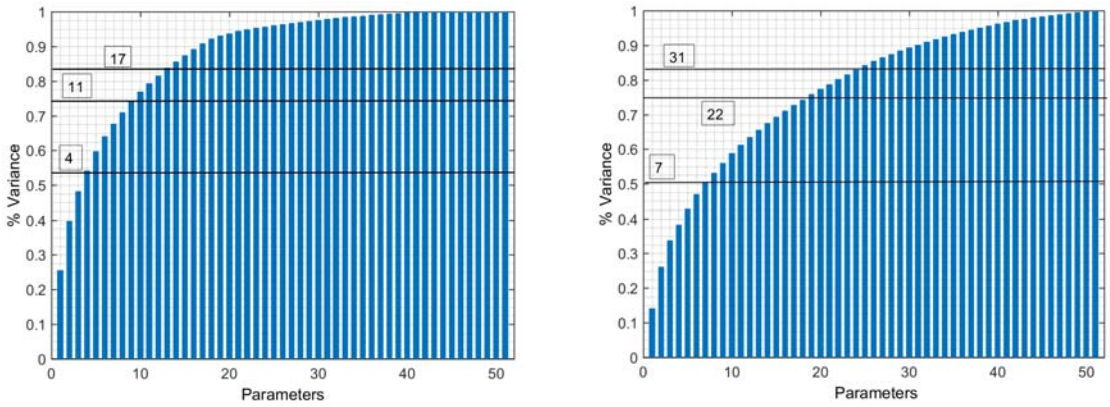


Figure 2. Sensitivity indices of (**a**) Antibody Concentration (**b**) Other Measured Metabolites Concentration to the parameters in the CHO cell model. In both Figure 2a,b, Parameters 1 to 4 correspond to the four process variables and parameter 5 to 51 correspond to the 47 kinetic parameters.

We plot the percentage variance explained of antibody and metabolite concentrations as a function of number of parameters in Figure 3a,b, respectively. The four most important parameters are two process variables, (#2 and #3), and two kinetic parameters (#20 and #41). Together, these account for 50% of the variance of antibody concentration, as shown in Figure 3a. We assign these parameters to subset A1. Subset A2 consists of seven parameters, including three process variables, #1-3, and four kinetic parameters, #13, #20, #41, and #46. Together, these 7 parameters account for 50% of the variance of the metabolite outputs, as shown in Figure 3b. We obtain the most important subset of parameters, subset A, by combining the kinetic parameters in A1 and A2. We arrive at subset A consisting of four parameters, #13, #20, #41, and #46. The process parameters, #1-3, are not included in this subset, because they are externally defined by the operating conditions and thus are not estimated from the experimental data. In a similar manner, 18 additional important parameters are identified and assigned to subset B. Together with the 4 most important parameters of subset A, the subset B parameters explain 80% of the variance of the antibody and metabolite concentrations. Seven additional parameters are

identified in subset C. The remaining 18 parameters explain the last 10% of the overall variance and are part of subset D.



(**a**) (**b**)

Figure 3. Accumulated sum of sensitivity indices: (**a**) Antibody Concentration (**b**) Other Measured Metabolites Concentration to 51 parameters in the CHO Model.

We first estimate the values of subset A parameters, while holding the other parameters fixed at their nominal values. The estimated values of these parameters in the first cycle of estimation are listed in Table 1, column 4. The parameters in the other subsets (B, C, and D) are at their nominal values as given in column 3. The Sum of Squared Error (SSE) between the predicted and measured outputs is 6.85 as listed in column 2 of Table 2. For comparison, we estimate all 47 parameters simultaneously using the same data. The SSE associated with the obtained model is 8.36, as given in column 5 of Table 2. This value is 22% larger than the SSE resulting from the estimation of only the 4 most important parameters in subset A. Moreover, the computational time for estimating all 47 parameters simultaneously is 5.48 h, whereas estimating the four parameters in subset A required only 0.48 h. All nonlinear parameter estimation tasks were performed in MATLAB using the *fmincon* function on a personal computer with 4 GB RAM memory and Intel Core i5-2500 (3.3 GHz) CPU. These results show that as the dimensionality of the parameter estimation problem is reduced, the accuracy of the model is improved, while the computational time is reduced. The SSE of the original model [4] is 11.68, which was obtained by estimating all of the model parameters simultaneously using simulated annealing. The difference between the SSEs of the models with the simultaneously estimated parameters (this study vs. [4]) suggests that when the dimensionality of the parameter estimation problem is large relative to the available data, the optimization may converge to different local minima depending on the algorithm.

Table 2. Comparison Sum of Squared Error of obtained models by sequentially and simultaneously re-estimating model parameters.

	Subset A	Subset A + B	Subset A + B + C	Simultaneous
Sum of Squared Error	6.85	6.63	6.60	8.36
Difference in SSE (%) [1]	0	−3.2	−3.7	22.0
Computational Time (h)	0.48	1.25	3.16	5.48

[1] The percentage difference uses SSE for subset A as the reference. A positive value means the corresponding SSE is larger than the SSE of subset A while negative value indicates a smaller SSE than the one of subset A.

By using the estimated values of the parameters in subset A in the first round as the initial guess, we estimate the parameters in subsets A and B in the next round. The initial values of the subset B parameters are their scaled values in column 3 of Table 1. The parameter values estimated in the second round are given in column 5 of Table 1. The corresponding SSE is 6.63 and it is given in column

3 of Table 2. With 18 additional parameters in subset B estimated, the SSE is only slightly (3.3%) smaller than the SSE obtained after the first round, while the computational time increases to 1.25 h from 0.48 h. This confirms that the parameters in subset B have a smaller impact on the model accuracy compared to subset A. With the best parameter values for subsets A and B as the initial guesses for the respective parameters, we estimate in the third round the 7 parameters in subset C, as well as the parameters in subsets A and B. After the third round, the SSE is further reduced by 0.5%. This very modest improvement in SSE again underscores that the most important parameters identified in subset A have the largest impact on model accuracy. Indeed, estimating just the 22 parameters in subsets A and B, while keeping the remaining 25 parameters fixed at their nominal values would have yielded a model that is just as accurate as the model obtained by estimating all parameters in subsets A, B, and C.

Taken together, the above results suggested that the sequential parameter estimation strategy could yield a more accurate model, while also reducing the computational time required for parameter estimation. To visually assess the accuracy of the sequentially estimated model, we plotted the model outputs obtained with the different parameter estimation strategies (Figure 4). For comparison, the experimental data used to estimate the original model [4] are also plotted in the same figure. The concentrations predicted by the model with simultaneously estimated parameters (Model 1) are shown in dashed lines, while the concentrations predicted by the model in which only the four most important parameters (subset A) are estimated (Model 2) are shown in solid line. Model 2 has a more accurate prediction in the concentrations of antibody (ANTI), the output of the greatest interest. In addition, Model 2 more accurately predicted the BIOM, GLC, ASP and SER concentration profiles.

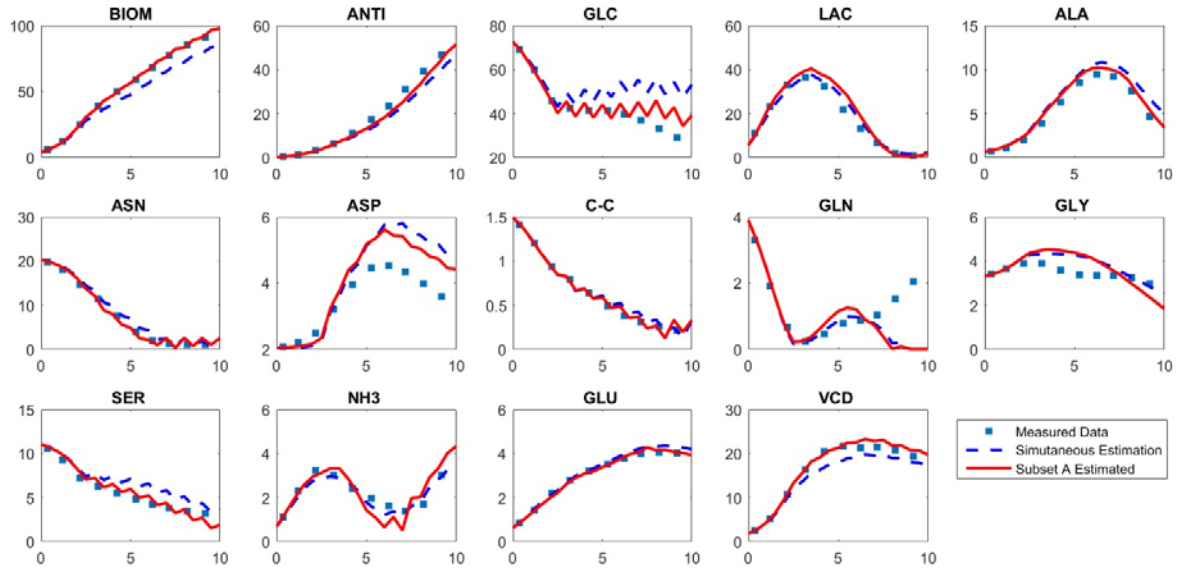


Figure 4. Comparison between predictions by the new model with 4 most sensitive parameters estimated (—) with predictions by the model with all parameters estimated simultaneously (–). The experimental data from [4] are shown, as well (■). The x-axis is the culture time in days and y-axis is the concentration in mM, except for the viable cell density (VCD) which is 10^6 cells/mL.

6. Conclusions

In this paper, we propose a sequential parameter estimation approach to improve the accuracy of the obtained model. The parameters to be estimated are first assigned to four subsets, A, B, C, and D, based on how sensitive the model predications are on the parameter values, quantified by their sensitivity indices. The parameters in subsets A, B and C correspond to the most important, important, and less important parameters, respectively. The least important parameters are grouped in subset D, and contribute only 10% to the model's outputs. The sensitivity indices are calculated using a

refined approach originating from the global sensitivity analysis via RSM model proposed by [18]. Instead of running Monte Carlo simulations on the estimated RSM models, we analytically calculate the sensitivity indices of each parameter. This further reduces the computational cost.

In the proposed sequential estimation of model parameters, one initially estimates the parameters in subset A, then in sets A and B. If enough data is available, the parameters in subsets A, B and C are then estimated. As shown in this paper, fitting the parameters in subset D will have negligible impact on the model's accuracy. When we estimated all the parameters simultaneously, including those in subset D, we obtained a statistically less accurate model, which confirms the efficacy of the proposed method. Avoidance of local minima in the related optimization task is conjectured to be the main reason for the superior performance of the sequential estimation of the model's parameters.

We demonstrate the benefits of the proposed method using a case study on a dynamic model of CHO cell metabolism in fed-batch culture. The model parameters are separated into four subsets: A, B, C, and D of deceasing importance. The first subset accounts for 50% of the outputs' variance, while subsets B and C account for an additional 30% and 10% variance, respectively. The corresponding SSE indicates that by estimating the very small subset of most important parameter (subset A), we can obtain an accurate starting model. If we then follow up with the estimation of the parameters in subsets A and B, an even more accurate model can be obtained. If additional parameters of less importance are estimated, the further improvement on model accuracy is minimal. If we estimate all of the parameters simultaneously, a less accurate model is achieved compared to the sequentially estimated model. We speculate that this might be due to the existence of several local minima, although additional work is warranted to more thoroughly explain this result. At least for the CHO model investigated in this paper, the observation that the simultaneously estimated model affords lower accuracy and highlights the potential benefit of the sequential estimation approach proposed here.

Author Contributions: Conceptualization, C.G. and K.L.; Methodology, C.G., Z.W.; Investigation, Z.W., C.G., H.S. and K.L.; Formal analysis, Z.W., C.G., H.S. and K.L.; Validation, K.L.; Software, Z.W., H.S.; Writing-Original Draft, Z.W., C.G. and K.L.; Writing-Review, H.S.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, W.; Kantardjieff, A. Mammalian Cell Cultures for Biologics Manufacturing. In *Mammalian Cell Cultures for Biologics Manufacturing*; Zhou, W., Kantardjieff, A., Eds.; Springer: Berlin/Heidelberger, Germany, 2014.
2. Nolan, R.P.; Lee, K. Dynamic model for CHO cell engineering. *J. Biotechnol.* **2012**, *158*, 24–33. [CrossRef] [PubMed]
3. Sanderson, C.S.; Barford, J.P.; Barton, G.W. A structured, dynamic model for animal cell culture systems. *Biochem. Eng. J.* **1999**, *3*, 203–211. [CrossRef]
4. Nolan, R.P.; Lee, K. Dynamic model of CHO cell metabolism. *Metab. Eng.* **2011**, *13*, 108–124. [CrossRef] [PubMed]
5. Mulukutla, B.C.; Gramer, M.; Hu, W.-S. On metabolic shift to lactate consumption in fed-batch culture of mammalian cells. *Metab. Eng.* **2012**, *14*, 138–149. [CrossRef] [PubMed]
6. Raue, A.; Kreutz, C.; Maiwald, T.; Bachmann, J.; Schilling, M.; Klingmüller, U.; Timmer, J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **2009**, *25*, 1923–1929. [CrossRef] [PubMed]
7. Kravaris, C.; Hahn, J.; Chu, Y. Advances and selected recent developments in state and parameter estimation. *Comput. Chem. Eng.* **2013**, *51*, 111–123. [CrossRef]
8. Saltelli, A.; Annoni, P. How to avoid a perfunctory sensitivity analysis. *Environ. Model. Softw.* **2010**, *25*, 1508–1517. [CrossRef]
9. Yao, K.Z.; Shaw, B.M.; Kou, B.; McAuley, K.B.; Bacon, D.W. Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design. *Polym. React. Eng.* **2003**, *11*, 563–588. [CrossRef]

10. Lee, D.; Ding, Y.; Jayaraman, A.; Kwon, J. Mathematical Modeling and Parameter Estimation of Intracellular Signaling Pathway: Application to LPS-induced NFκB Activation and TNFα Production in Macrophages. *Processes* **2018**, *6*, 21. [CrossRef]
11. McLean, K.A.P.; Wu, S.; McAuley, K.B. Mean-Squared-Error Methods for Selecting Optimal Parameter Subsets for Estimation. *Ind. Eng. Chem. Res.* **2012**, *51*, 6105–6115. [CrossRef]
12. Eghtesadi, Z.; McAuley, K.B. Mean-squared-error-based method for parameter ranking and selection with noninvertible fisher information matrix. *AIChE J.* **2016**, *62*, 1112–1125. [CrossRef]
13. Degenring, D.; Froemel, C.; Dikta, G.; Takors, R. Sensitivity analysis for the reduction of complex metabolism models. *J. Process Control* **2004**, *14*, 729–745. [CrossRef]
14. Sobol, I.M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **2001**, *55*, 271–280. [CrossRef]
15. Ho, Y.; Varley, J.; Mantalaris, A. Development and analysis of a mathematical model for antibody-producing GS-NS0 cells under normal and hyperosmotic culture conditions. *Biotechnol. Prog.* **2006**, *22*, 1560–1569. [CrossRef] [PubMed]
16. Zheng, Y.; Rundell, A. Comparative study of parameter sensitivity analyses of the TCR-activated Erk-MAPK signalling pathway. *IEE Proc. Syst. Biol.* **2006**, *153*, 201–211. [CrossRef]
17. Mailier, J.; Delmotte, A.; Cloutier, M.; Jolicoeur, M.; Wouwer, A.V. Parametric Sensitivity Analysis and Reduction of a Detailed Nutritional Model of Plant Cell Cultures. *Biotechnol. Bioeng.* **2011**, *108*, 1108–1118. [CrossRef] [PubMed]
18. Kiparissides, A.; Georgakis, C.; Mantalaris, A.; Pistikopoulos, E.N. Design of In Silico Experiments as a Tool for Nonlinear Sensitivity Analysis of Knowledge-Driven Models. *Ind. Eng. Chem. Res.* **2014**, *53*, 7517–7525. [CrossRef]
19. Montgomery, D.C. *Design and Analysis of Experiments*, 8th ed.; Wiley: New York, NY, USA, 2013.
20. Saltelli, A. Making best use of model evaluations to compute sensitivity indices. *Comput. Phys. Commun.* **2002**, *145*, 280–297. [CrossRef]
21. Draper, N.R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, NY, USA, 1998.
22. MathWorks. *Optimization Toolbox™ User's Guide (2015b)*; MathWorks Inc.: Natick, MA, USA, 2015.
23. Byrd, R.H.; Hribar, M.E.; Nocedal, J. An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM J. Optim.* **1999**, *9*, 877–900. [CrossRef]