Chemie Ingenieur Technik

On the use of online reparametrization in automated platforms for kinetic model identification

Journal:	*Chemie Ingenieur Technik*
Manuscript ID	cite.201800095.R1
Wiley - Manuscript type:	Forschungsarbeit
Date Submitted by the Author:	n/a
Complete List of Authors:	Quaglio, Marco; University College London, Department of Chemical Engineering Waldron, Conor; University College London, Department of Chemical Engineering Pankajakshan, Arun; University College London, Department of Chemical Engineering Cao, Enhong; University College London, Department of Chemical Engineering Gavriilidis, Asterios; University College London, Chemical Engineering Dept Fraga, Eric; University College London, Department of Chemical Engineering Galvanin, Federico; University College London, Department of Chemical Engineering
Keywords:	design of experiments, identification, model, online, robust parametrization

On the use of online reparametrization in automated platforms for kinetic model identification

Marco Quaglio, Conor Waldron, Arun Pankajakshan, Enhong Cao, Asterios Gavriilidis, Eric S. Fraga, Federico Galvanin[1]

Abstract

Parameter estimation algorithms integrated in automated platforms for kinetic model identification are required to solve two optimization problems: i) a parameter estimation problem given the available samples; ii) a model-based design of experiments problem to select the conditions for collecting future samples. These problems may be ill-posed, leading to numerical failures when optimization routines are applied. In this work, an approach of online reparametrization is introduced to enhance the robustness of model identification algorithms towards ill-posed parameter estimation problems.

keywords: design of experiments, identification, model, online, robust parametrization

1. Introduction

Automated model identification platforms were recently employed to perform unmanned experimental campaigns with the aim of collecting data for estimating the parameters of kinetic models [1–3]. These devices have the potential to dramatically speed up the study of kinetic phenomena and reduce the cost of the experimental activity. The model identification algorithms implemented in these platforms integrate two computational tools: 1) a tool for model-based design of experiments (MBDoE) to select the optimal conditions for the collection of future samples with the aim of improving the statistical quality of the parameter estimates [4] and 2) a tool for computing parameter estimates given the samples collected by the automated system [5]. Both these tools need to solve optimization problems, for which numerical optimization routines are required. The effectiveness of model identification algorithms requires the objective functions of the aforementioned optimization problems to be well-posed [6].

The problem of estimating the parameters of kinetic models is frequently ill-conditioned [7]. Identifiability problems occur whenever the fitted model responses are poorly sensitive to a change of some parameters and/or extreme correlation among parameter pairs is present. Under these circumstances, the model is called sloppy (also called poorly constrained model) [7]. In the presence of a sloppy model, the objective functions of both parameter estimation and MBDoE problems may be ill-conditioned, resulting in significant numerical failures in the process of model identification and a misuse of experimental resources. Enhancing the robustness of automated model identification platforms towards model sloppiness is necessary to promote their further diffusion into research laboratories. The ill-posedness of an optimization problem is quantified by its condition number, which is defined as the ratio between the largest and the smallest eigenvalues of the Hessian. The condition number summarizes in a scalar quantity the discrepancy in the sensitivities of the model responses towards a change in the model parameters and parameter correlations. Model identification algorithms are prone to numerical failures whenever the condition number is high [8].

A number of regularization techniques were proposed to solve ill-conditioned parameter estimation and experimental design problems [9–13]. Regularization involves the introduction of a bias in the parameter estimates with the aim of reducing their variance and, concomitantly, reducing the condition number of the

[1] Marco Quaglio, Conor Waldron, Arun Pankajakshan, Dr. Enhong Cao, Prof. Asterios Gavriilidis, Prof. Eric S. Fraga, Dr. Federico Galvanin; corresponding author: Dr. Federico Galvanin; e-mail: f.galvanin@ucl.ac.uk; Department of Chemical Engineering, University College London (UCL), Torrington Place, WC1E 7JE London, United Kingdom

problem [10]. Popular regularization techniques are i) the Tikhonov regularization [12, 13], ii) the truncated singular value decomposition [9, 12] and iii) the parameter subset selection [9, 11, 12]. Other studies recommend the use of reparametrization (RP) to address the practical identifiability problem of sloppy models [14–19]. Conversely to regularization, RP-based methods do not require the introduction of a bias in the parameter estimates. Instead, the aim of RP-based approaches is the transformation of the original parameter space into a more robust space, where the condition number is smaller and optimization algorithms can be applied more effectively. Tailored transformations have been suggested to reparametrize specific kinetic model structures, e.g. Arrhenius-type reaction rates [16–19]. However, only few systematic RP-based approaches are available in the literature [15]. A weakness of current RP-based approaches is that whenever a robust, i.e. non-sloppy, structure is identified for a given kinetic model, the parametrization is fixed until the end of the model identification process. However, sloppiness is associated with both the mathematical structure of the model equations and the dataset available to identify the model. It is not possible to guarantee that a robust model will not become sloppy after the collection and fitting of new samples [6].

In this manuscript, an online RP-based approach is introduced with the aim of enhancing the robustness of model identification algorithms towards model sloppiness. It is shown that the approach is particularly suited for the integration in automated platforms for online model identification. In the proposed framework, the online RP method automatically modifies the model parametrization after the collection of each sample with the aim of maintaining a small condition number throughout the whole experimental campaign. Online RP is demonstrated on a simulated case study where the aim is to estimate the kinetic parameters of an esterification reaction of benzoic acid with ethanol.

2. Methods

2.1. Proposed methodology

An automated platform for online model identification is available for studying the kinetics of a physical system of interest. A set $\mathbf{y}$ of N_y physical quantities can be sampled online by a measurement system. The measurement error for $\mathbf{y}$ is Gaussian with zero mean and known covariance matrix $\boldsymbol{\Sigma}$. The kinetic model in Eq. (1) is available to the scientist to describe the behavior of the physical system.

$$\begin{aligned} \mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t, \boldsymbol{\theta}) &= \mathbf{0} \\ \hat{\mathbf{y}} &= \mathbf{h}(\mathbf{x}) \end{aligned} \tag{1}$$

In Eq. (1), $\hat{\mathbf{y}}$ is a $N_y \times 1$ array of predictions for the measurable system states, $\mathbf{f}$ and $\mathbf{h}$ are respectively a $N_f \times 1$ and a $N_y \times 1$ set of functions, $\mathbf{x}$ is a $N_x \times 1$ set of state variables, $\mathbf{u}$ is a $N_u \times 1$ array of controllable system inputs, t is time and $\boldsymbol{\theta} \in \Theta$ is a $N_\theta \times 1$ array of parameters $\theta_1, \dots, \theta_{N_\theta}$. The objective of the scientist is estimating the set of parameters $\boldsymbol{\theta}$ with the highest possible precision by performing an experimental campaign on the automated platform. It is assumed that, in principle, the model parameters can be uniquely retrieved by fitting measurements of $\mathbf{y}$, i.e. the model satisfies the requirements for structural identifiability [20]. In other words, it is assumed that given sufficient measurements of $\mathbf{y}$, the fitting cost function admits a unique global optimum. In practice, the global optimizer of the cost function has to be identified employing numerical optimization routines and the convergence towards the optimal parameter values may be impractical if the model is sloppy [8].

A framework for parameter estimation implementing a step of online model reparametrization is now introduced with the aim of improving the robustness of online model identification algorithms towards model sloppiness. A diagram showing the proposed procedure is given in Fig. 1. The procedure starts from a preliminary set of N samples of $\mathbf{y}$, i.e. $Y = [\mathbf{y}_1, \dots, \mathbf{y}_N]$ and the kinetic model in Eq. (1). The set of equations in Eq. (1) is initially extended including the linear system of equations in Eq. (2).

$$\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\omega} \tag{2}$$

In Eq. (2), $\boldsymbol{\omega} \in \Omega$ is the $N_\theta \times 1$ array of parameters in a transformed parameter space Ω, $\mathbf{G}$ is a linear transformation from the transformed space Ω to the original parameter space Θ. At the beginning of the model identification process, the transformation matrix $\mathbf{G}$ is the identity matrix $\mathbf{I}$. In other words, the transformed and the original parameter spaces Ω and Θ are initially coincident. The available dataset is provided as input to the model identification algorithm. The fundamental steps in the algorithm are:

1. *A reparametrization step.* At this stage, model sloppiness is diagnosed by analyzing the Hessian of the log-likelihood function of the model. An update for the transformation $\mathbf{G}$ is then computed with the aim of eliminating the sloppiness (i.e. minimizing the condition number to unity) given the available dataset.

2. *A parameter estimation step.* The parameters $\boldsymbol{\omega} \in \Omega$ are estimated after the reparametrization step and their covariance matrix $\mathbf{V}_\omega$ is computed to quantify the statistical quality of the estimate. Estimates and covariance computed in Ω are then transformed to the original space Θ by applying the transformation $\mathbf{G}$ and returned to the user as output.

3. *An optimal MBDoE step.* If parameter statistics are not satisfactory and the experimental budget allows for the collection of additional samples, the experimental activity may continue. Optimal MBDoE methods are used at this stage to select optimal experimental conditions with the aim of minimizing the uncertainty on the parameter estimates [4]. Optimal conditions are then transmitted to the automated system for the collection of the following sample. In the proposed procedure the optimal MBDoE problem is solved in Ω with the aim of reducing the uncertainty on the estimates in the original space Θ.

In the proposed procedure, parameter estimation and the optimal MBDoE problems are solved calling optimization routines in a conveniently transformed parameter space Ω. Estimates are then transformed to the original parameter space Θ by applying algebraic transformations. The three steps of the procedure are further detailed in the following subsections.

((Figure 1))

INSERT FIGURE 1 {methodology.tif}

2.1.1. Reparametrization

Let $\mathbf{G}_\mathrm{P}$ be the *primary* transformation matrix before the reparametrization stage in the procedure. $\mathbf{G}_\mathrm{P}$ is initially the identity matrix $\mathbf{I}$. Let $\mathbf{H}|_{\mathbf{G}=\mathbf{G}_\mathrm{P}}$ be the negative Hessian of the log-likelihood function $\Phi(\boldsymbol{\omega}|Y)|_{\mathbf{G}=\mathbf{G}_\mathrm{P}}$ computed with $\mathbf{G} = \mathbf{G}_\mathrm{P}$. The negative Hessian of the log-likelihood function is also called the observed Fisher information matrix and its inverse provides a quantification of the covariance matrix of the parameter estimates [21].

The eigendecomposition of $\mathbf{H}|_{\mathbf{G}=\mathbf{G}_\mathrm{P}}$ is performed to diagnose the geometry of the parameter space and quantify the sloppiness of the model given the available dataset Y. Let $\lambda_1, \dots, \lambda_{N_\theta}$ be the eigenvalues of $\mathbf{H}|_{\mathbf{G}=\mathbf{G}_\mathrm{P}}$ and let $\boldsymbol{\Lambda}$ be the diagonal matrix whose ii-th element is λ_i. The ratio between the maximum and the minimum eigenvalue represents the condition number κ.

$$\kappa = \frac{\max \lambda_i}{\min \lambda_i} \tag{3}$$

Let matrix $\mathbf{U}$ be the orthonormal basis of right eigenvectors of $\mathbf{H}|_{\mathbf{G}=\mathbf{G}_\mathrm{P}}$. Matrices $\mathbf{\Lambda}$ and $\mathbf{U}$ respectively quantify the extent of the sloppiness and the directions of the parameter space that are associated to the sloppiness [9]. A family of *secondary* transformation updates $\mathbf{G}_\mathrm{S}$ is built from $\mathbf{G}_\mathrm{P}$, $\mathbf{U}$ and $\mathbf{\Lambda}$ as in Eq. (4) for minimizing the condition number of the problem (i.e. making $\kappa = 1.0$).

$$\mathbf{G}_\mathrm{S} = d\mathbf{G}_\mathrm{P}\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{R} \tag{4}$$

The family of transformations in Eq. (4) includes a scaling factor $d \neq 0$ and a rotation matrix in the parameter space $\mathbf{R}$. The condition number κ is not influenced by the choice of d and $\mathbf{R}$. However, d and $\mathbf{R}$ can be regarded as scaling factors for the parameters. Model identification algorithms may be influenced by the relative scale of parameter estimates, e.g. in the computation of gradients and parameter sensitivities [22]. In this work, d and $\mathbf{R}$ are computed to scale the transformed model parameters to the same order of magnitude (more information on the computation of d and $\mathbf{R}$ can be found in the supporting information). The *primary* transformation $\mathbf{G}_\mathrm{P}$ is then *updated*, i.e. $\mathbf{G}_\mathrm{P}$ is replaced with $\mathbf{G}_\mathrm{S}$, for the following call of the model identification algorithm.

2.1.2. Parameter estimation

Parameter estimation is performed by solving an optimization problem in the transformed parameter space Ω. The transformation matrix $\mathbf{G}$ is set equal to $\mathbf{G}_\mathrm{S}$. The log-likelihood function $\Phi(\boldsymbol{\omega}|Y)|_{\mathbf{G}=\mathbf{G}_\mathrm{S}}$ is then computed and optimized as in Eq. (5) to obtain the maximum likelihood estimate $\hat{\boldsymbol{\omega}}$.

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}\in\Omega} \Phi(\boldsymbol{\omega}|Y)|_{\mathbf{G}=\mathbf{G}_\mathrm{S}} \tag{5}$$

The covariance $\mathbf{V}_\omega$ for the estimates $\hat{\boldsymbol{\omega}}$ is then computed as in Eq. (6) by inverting the observed Fisher information matrix $\mathbf{H}|_{\mathbf{G}=\mathbf{G}_\mathrm{S}}$ [5].

$$\mathbf{V}_\omega = \left[\mathbf{H}|_{\mathbf{G}=\mathbf{G}_\mathrm{S}}\right]^{-1} \tag{6}$$

The parameter estimates $\hat{\boldsymbol{\omega}}$ and related covariance $\mathbf{V}_\omega$ computed in the transformed space Ω are then transformed to the original parameter space Θ by applying the transformation $\mathbf{G}_\mathrm{S}$. The operation leads to the computation of the maximum likelihood estimate $\boldsymbol{\theta} \in \Theta$ and its associated covariance matrix $\mathbf{V}_\theta$.

$$\hat{\boldsymbol{\theta}} = \mathbf{G}_\mathrm{S}\hat{\boldsymbol{\omega}} \tag{7}$$

$$\mathbf{V}_\theta = \mathbf{G}_\mathrm{S}\mathbf{V}_\omega\mathbf{G}_\mathrm{S}^T \tag{8}$$

Confidence intervals for the estimates $\hat{\boldsymbol{\theta}}$ and correlation coefficients c_{ij} between any parameter pair θ_1 and θ_2 can be directly computed from $\mathbf{V}_\theta$ [5]. The expression used to compute the correlation coefficients c_{ij} is given in Eq. (9), where quantity $v_{\theta,ij}$ represents the ij-th element of $\mathbf{V}_\theta$.

$$c_{ij} = \frac{v_{\theta,ij}}{\sqrt{v_{\theta,ii}v_{\theta,jj}}} \quad \forall\, i,j \tag{9}$$

2.1.3. Optimal MBDoE for parameter precision

The inversion of an ill-conditioned matrix may be required to solve the MBDoE problem if the model is sloppy [7, 9]. Hence, it is proposed to solve the MBDoE problem in the robust space Ω with the aim of minimizing parameter uncertainty in the original space Θ. A class of popular design criteria is represented by the so-called *alphabetic* criteria, namely A-optimal, D-optimal and E-optimal [4, 21, 23]. In general, optimal MBDoE conditions are sensitive to the model parametrization and to the choice of the design criterion. In this study, only the D-optimal criterion is used as it is invariant under linear transformations of the parameter space [24, 25].

A prediction for the parameter covariance $\widehat{\mathbf{V}}_\omega$ (i.e. the posterior covariance in Ω) after the collection of the sample to be designed is computed as in Eq. (10).

$$\widehat{\mathbf{V}}_\omega = \left[\mathbf{V}_\omega^{-1} + \nabla\hat{\mathbf{y}}(\widehat{\boldsymbol{\omega}})\boldsymbol{\Sigma}^{-1}\nabla\hat{\mathbf{y}}(\widehat{\boldsymbol{\omega}})^T|_{\mathbf{G}=\mathbf{G}_S}\right]^{-1} \tag{10}$$

In Eq. (10), ∇ is the gradient operator in the parameter space; the inverse of the prior covariance $\mathbf{V}_\omega^{-1}$ is included to quantify the information associated to previously fitted samples; the second addend in the bracket quantifies the expected information of the sample to be designed as a function of the design vector $\boldsymbol{\varphi} = [\mathbf{u}, t_s]$, where $\mathbf{u} \in U$ is the array of model inputs and $t_s \in [t_{\mathrm{MIN}}, t_{\mathrm{MAX}}]$ is the sampling time. The D-optimal conditions $\boldsymbol{\varphi}^* = [\mathbf{u}^*, t_s^*]$ are computed by minimizing the determinant of the predicted covariance.

$$\begin{aligned} \boldsymbol{\varphi}^* &= \arg\min_{\boldsymbol{\varphi}} \det \widehat{\mathbf{V}}_\omega \\ \text{s.t. } \mathbf{u} &\in U,\ t_s \in [t_{\mathrm{MIN}}, t_{\mathrm{MAX}}] \end{aligned} \tag{11}$$

The optimized conditions $\mathbf{u}^*$, t_s^* are transmitted to the automated setup for the collection of the new sample (see Fig. 1).

3. Case study

The proposed approach to online reparametrization (i.e. online RP) is tested on a simulated case study where the objective is the identification of a kinetic model of esterification of benzoic acid and ethanol in flow reactor. The kinetic model is presented in Sect. 3.1 and the methods adopted for testing the methodology are presented in Sect. 3.2.

3.1. Kinetic model

The kinetic mechanism is modelled as a single reaction where benzoic acid (BA) and ethanol (Et) react producing ethyl benzoate (EB) and water (W) [26].

$$\mathrm{BA+Et} \rightarrow \mathrm{EB+W} \tag{12}$$

The reaction is assumed to occur in an ideal plug-flow reactor operated at steady-state, isothermal conditions. The reactor length is assumed to be 2m. The reaction kinetics is modelled as a first order in the benzoic acid. These assumptions are translated into the set of ordinary differential equations in Eq. (13).

$$v\frac{dC_i}{dz} = \nu_i k C_{\mathrm{BA}} \quad \forall\, i = \mathrm{BA, Et, EB, W} \tag{13}$$

In Eq. (13), C_i [mol L^{-1}] represents the concentration of the i-th species; z [m] is the axial coordinate of the flow reactor; v [m s^{-1}] is the axial velocity of the liquid mixture; ν_i is the stoichiometric coefficient of species i; k [s^{-1}] is an Arrhenius-type rate constant. The following parametrization with two parameters $\boldsymbol{\theta} = [\theta_1, \theta_2]$ is assumed for k:

$$k = e^{\theta_1 - \frac{\theta_2 10^4}{RT}} \tag{14}$$

In Eq. (14), T [K] is the temperature and R [J mol^{-1} K^{-1}] is the ideal gas constant. The use of a parametrization in the form of Eq. (14) generally reduces the condition number of the problem with respect to the original structure of the Arrhenius constant, i.e. $k = Ae^{-E_a/RT}$, where the parameters are a pre-exponential factor A [s^{-1}] and an activation energy E_a [J mol^{-1}] [16, 19].

3.2. Objective and methods

The objective in this case study is the estimation of the kinetic parameter set $\boldsymbol{\theta} = [\theta_1, \theta_2]$. A model identification algorithm reflecting the framework presented in Sect. 2.1 was implemented in Python 2.7. An option was included in the algorithm to activate or deactivate the *reparametrization step* in the procedure for testing the algorithm both in the absence and in the presence of online RP. It is assumed that the experimental budget allows for the collection of 8 samples. A sample is constituted by the two measurements of benzoic acid and ethyl benzoate concentration at the outlet of the reactor, i.e. $\mathbf{y} = \left[C_{\mathrm{BA}}^{\mathrm{OUT}}, C_{\mathrm{EB}}^{\mathrm{OUT}}\right]$. The measurement noise associated to the sample is described by the covariance $\boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} 9.0 \cdot 10^{-4} & 0.0 \\ 0.0 & 2.5 \cdot 10^{-5} \end{pmatrix} \tag{15}$$

The design space for the collection of the samples is assumed as three-dimensional. The independent experimental conditions are the inlet concentration of benzoic acid $C_{\mathrm{BA}}^{\mathrm{IN}}$ in the range $0.9 - 1.55$ mol L^{-1}, the flowrate F in the range $7.5 - 30.0$ µL min^{-1}; the temperature T in the range $343.0 - 413.0$ K. The conditions for the collection of sample 1 and sample 2 are pre-defined. Sample 1 is collected setting $C_{\mathrm{BA}}^{\mathrm{IN}} = 1.50$ mol L^{-1}, $F = 20.0$ µL min^{-1} and $T = 413.0$ K. Sample 2 is collected setting $C_{\mathrm{BA}}^{\mathrm{IN}} = 1.00$ mol L^{-1}, $F = 10.0$ µL min^{-1} and $T = 393.0$ K. Samples 3 to 8 are designed by the model identification algorithm adopting a D-optimal criterion (see Sect. 2.1.3). Samples are generated *in-silico* by integrating the kinetic model in Eq. (13) and setting the parameters equal to the value $\boldsymbol{\theta}^* = [15.27, 7.60]$ (notice that setting $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ corresponds to setting the pre-exponential factor $A = e^{\theta_1} = 4.3 \cdot 10^6$ s^{-1} and the activation energy $E_a = \theta_2 \cdot 10^4 = 7.6 \cdot 10^4$ J mol^{-1}. Gaussian noise with covariance $\boldsymbol{\Sigma}$ is added to the simulation results to simulate the measurement error.

Two campaigns are simulated to test the proposed online reparametrization approach:

1. *Non-RP campaign.* In the non-RP campaign the online RP in not active.

2. *RP campaign.* In the RP campaign the online RP is active.

The two campaigns are performed to assess the influence of the online RP on the model identification while all the other components of the model identification algorithm are fixed. The benefit of the online RP is quantified performing a statistical test to compare the estimates $\widehat{\boldsymbol{\theta}}$ obtained in the RP and in the non-RP campaigns with the target parameter values $\boldsymbol{\theta}^* = [15.27, 7.60]$. This involves testing the null hypothesis that the following statistic χ_θ^2 is distributed as a χ^2 distribution with degree of freedom $N_\theta = 2$.

$$\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^T \mathbf{V}_\theta^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) = \chi_\theta^2 \sim \chi^2 \tag{16}$$

A small p-value of the statistic χ_θ^2 (e.g. smaller than $1.0\ \%$) is interpreted as an index of failure of the model identification algorithm in estimating the target parameter values.

4. Results and discussion

((Table 1))

((Table 2))

((Figure 2))
INSERT FIGURE 2 {condition_number.tif}

((Figure 3a)) ((Figure 3b))
INSERT FIGURE 3a {estimates.tif} AND FIGURE 3b {confidence_intervals.tif}

((Figure 4))
INSERT FIGURE 4 {final_parameter_statistics.tif}

Parameter estimates for θ_1 and θ_2 in the non-RP campaign are reported in Tab. 1 together with their respective 95% confidence intervals and coefficient of correlation c_{12}. Parameter correlation in Θ is above 99.97% throughout the non-RP campaign. In the non-RP campaign, optimization routines for parameter estimation and MBDoE are applied on the original parameter space Θ. The condition number κ in Θ remains above $7.2 \cdot 10^3$ throughout the campaign. The condition number κ is finite, suggesting that there exists a unique optimum for the log-likelihood function in the proximity of the estimates, i.e. the kinetic model is structurally identifiable [20]. However, a p-value of 0.0% after the collection of 8 samples suggests that the model identification algorithm produced estimates that are statistically inconsistent with the target parameter values. The failure of the numerical model identification algorithm in retrieving the target parameters is interpreted as a consequence of practical identifiability issues associated with the high condition number κ.

Parameter estimates and related statistics for the RP campaign are reported in Tab. 2. Also in the RP campaign, the correlation between θ_1 and θ_2 in Θ remains extremely high, i.e. above 99.85% throughout the whole experimental campaign. However, in the RP campaign, parameter estimation and MBDoE problems are solved by optimization routines in the transformed parameter space Ω. The condition number in Ω was computed at each *parameter estimation step* (i.e. after the *reparametrization step*) in the course of the RP campaign and it is reported in Tab. 2. As one can see from Tab. 2, the online RP iteratively reduces the condition number from the initial value $\kappa = 7.0 \cdot 10^5$ to the minimum value $\kappa = 1.0$ after five calls of the algorithm (i.e. after the collection of 6 samples). A final p-value of 78.13% is interpreted as an index of success of the RP campaign in retrieving the target parameters $\boldsymbol{\theta}^*$.

The condition number κ after the collection of each sample can be appreciated in Fig. 2 for both the non-RP and the RP campaign. In the RP case, the condition number is minimized to $\kappa = 1.0$ after the collection of 6 samples, i.e. 5 updates of the transformation matrix $\mathbf{G}$ are required to minimize the condition number. This is explained by the computation of an inappropriate update for the transformation matrix $\mathbf{G}$ at the first iterations of the algorithm. The update for the transformation matrix is computed as a function of the Hessian $\mathbf{H}$ computed setting $\mathbf{G}$ equal to the primary transformation matrix $\mathbf{G}_\mathrm{P}$ (see Sect. 2.1). The condition number κ in the space associated to $\mathbf{G}_\mathrm{P}$ may be extremely high at the first iteration of the algorithm. An initially high condition number may lead to an inaccurate computation of the Hessian and, consequently, lead to the

computation of an inappropriate update for the transformation matrix $\mathbf{G}$. However, as one can see from Fig. 2, the condition number is reduced to the minimum value $\kappa = 1.0$ in few iterations.

The estimated values of model parameters and respective confidence intervals reported in Tab. 1 and Tab. 2 are shown, respectively, in Fig. 3a and Fig. 3b for both the non-RP case (dotted line) and for the RP case (solid line). Results clearly show how the reparametrization affects both the accuracy of the estimates (closeness to the reference value) (see Fig. 3a), with the RP-case providing a faster convergence to the true value of model parameters in this case, and the precision of the estimates, with confidence intervals which are different (see Fig. 3b). In particular, the confidence interval for parameter θ_1, although very small, is higher in the RP-case than in the non-RP case. The 95% confidence ellipsoids associated to the final estimates in the RP and in the non-RP campaign are compared in Fig. 4. The target parameter value $\boldsymbol{\theta}^*$ is represented in the figure as a star-shaped symbol. As one can see from Fig. 4, the ellipsoid computed in the non-RP campaign (dotted ellipsoid) does not contain the target parameters while the ellipsoid computed in the RP campaign (solid ellipsoid) contains the target parameters. The graph shows that the non-RP campaign, performed without online RP, would lead to the conclusion that the target parameters *are not* the kinetic constants of the system. The RP campaign led to a more *robust* estimation of the target kinetic constants.

Monotonic convergence of the estimates to the target parameter values is recognized as an important index to assess the effectiveness of regularization methods implemented in parameter estimation algorithms [10]. Nonetheless, it is recognized that the robustness of RP-based methods relies on the monotonic decrease of the condition number in the course of the experimental campaign. Improving the proposed approach to ensure *smooth* (i.e. monotonic) convergence of the condition number to 1.0 is going to be focus of future research activities.

5. Conclusion

A model identification algorithm implementing a novel approach of online reparametrization, i.e. an approach of online transformation of the model parameter space, was proposed in this manuscript. The approach was designed specifically to deal with the problem of parameter estimation in the presence of *sloppy* model structures. In other words, the method was developed for situations where, in principle, model parameters can be uniquely retrieved fitting experimental data (i.e. the model satisfies the requirements for structural identifiability), but the small sensitivity of the model responses to a parameter change and/or high parameter correlation hinders the numerical estimation of the parameter values.

The proposed algorithm was tested on a simulated case study for estimating the kinetic parameters in a two-parameters model of benzoic acid and ethanol esterification in a flow reactor. The presented algorithm iteratively reduced and eventually eliminated model sloppiness minimizing the condition number of an originally ill-conditioned model identification problem. It was possible to statistically demonstrate that the sloppy nature of the model was preventing a conventional model identification algorithm from retrieving the target value of the kinetic constants. The presented model identification algorithm was instead capable of computing estimates that were statistically consistent with the assumed target values of the kinetic parameters.

Future research activities will focus on three aspects: 1) integrating of the proposed online RP method in an experimental automated model identification platform; 2) improving the efficiency of the proposed method reducing the number of iterations required to minimize the condition number; 3) validating the proposed approach on more complex model structures, e.g. kinetic models involving more than two parameters and more measured responses.

6. Acknowledgements

Financial support for the work presented in this manuscript was offered by: Hugh Walter Stern PhD Scholarship, University College London; EPSRC Grant EP/J017868/1.

7. Symbols used

Symbols – scalars

A	$[\mathrm{s}^{-1}]$ pre-exponential factor
c_{ij}	$[-]$ correlation coefficient for i-th and j-th parameter in $\widehat{\boldsymbol{\theta}}$
C_i	$[\mathrm{mol\ L}^{-1}]$ concentration of species i
C_i^{IN}	$[\mathrm{mol\ L}^{-1}]$ concentration of species i at the inlet
C_i^{OUT}	$[\mathrm{mol\ L}^{-1}]$ concentration of species i at the outlet
d	$[-]$ scaling factor of parameter space ($\neq 0$)
E_a	$[\mathrm{J\ mol}^{-1}]$ activation energy
F	$[\mu\mathrm{L\ min}^{-1}]$ volumetric flowrate
k	$[\mathrm{s}^{-1}]$ kinetic constant
N	$[-]$ number of samples in the available dataset Y
N_f	$[-]$ number of functions in the given kinetic model
N_u	$[-]$ number of independent inputs in a given kinetic model
N_x	$[-]$ number of state variables in a given kinetic model
N_y	$[-]$ number of output variables in a given kinetic model
N_θ	$[-]$ number of non-measurable parameters in a given model
R	$[\mathrm{J\ mol}^{-1}\ \mathrm{K}^{-1}]$ ideal gas constant
t	$[\mathrm{s}]$ time
t_s	$[\mathrm{s}]$ sampling time
t_s^*	$[\mathrm{s}]$ D-optimal sampling time
t_{MIN}	$[\mathrm{s}]$ lower bound for the sampling time
t_{MAX}	$[\mathrm{s}]$ upper bound for the sampling time
T	$[\mathrm{K}]$ temperature
v	$[\mathrm{m\ s}^{-1}]$ flow velocity along the axial coordinate of flow reactor
$v_{\theta,ij}$	[various units] ij-th element of covariance matrix $\mathbf{V}_{\boldsymbol{\theta}}$
Y	$[N]$ dataset available for model identification
z	$[\mathrm{m}]$ axial coordinate of flow reactor

Symbols - vectors and matrices

$\mathbf{f}$	$[N_f \times 1]$ column array of functions
$\mathbf{G}$	$[N_\theta \times N_\theta]$ linear transformation of parameter space $\Omega \to \Theta$
$\mathbf{G}_{\mathrm{P}}$	$[N_\theta \times N_\theta]$ primary transformation of parameter space $\Omega \to \Theta$
$\mathbf{G}_{\mathrm{S}}$	$[N_\theta \times N_\theta]$ secondary transformation of parameter space $\Omega \to \Theta$
$\mathbf{h}$	$[N_y \times 1]$ column array of functions for the model output variables
$\mathbf{H}$	$[N_\theta \times N_\theta]$ observed Fisher information matrix
$\mathbf{I}$	$[N_\theta \times N_\theta]$ identity matrix
$\mathbf{R}$	$[N_\theta \times N_\theta]$ matrix of rotation of parameter space
$\mathbf{u}$	$[N_u \times 1]$ column array of independent control variables (model inputs)
$\mathbf{u}^*$	$[N_u \times 1]$ D-optimal values for control variables (model inputs)
$\mathbf{U}$	$[N_\theta \times N_\theta]$ right normalized eigenbasis of $\mathbf{H}$
$\mathbf{V}_\theta$	$[N_\theta \times N_\theta]$ covariance of parameter estimates in Θ
$\mathbf{V}_\omega$	$[N_\theta \times N_\theta]$ covariance of parameter estimates in Ω
$\widehat{\mathbf{V}}_\theta$	$[N_\theta \times N_\theta]$ predicted covariance of parameter estimates in Ω
$\mathbf{x}$	$[N_x \times 1]$ column array of state variables
$\mathbf{y}$	$[N_y \times 1]$ sample - column array of measured output variables
$\mathbf{y}_i$	$[N_y \times 1]$ i-th sample in dataset Y
$\hat{\mathbf{y}}$	$[N_y \times 1]$ column array of predicted output variables

Greek symbols – scalars

θ_i	[various units] i-th model parameter

Θ	$[N_\theta]$ original vector space of model parameters
κ	$[-]$ condition number
λ_i	$[-]$ i-th eigenvalue of $\mathbf{H}$
ν_i	$[-]$ stoichiometric of the i-th species
Φ	$[-]$ log-likelihood function
Ω	$[N_\theta]$ transformed vector space of model parameters
χ^2	$[-]$ denotes a chi-squared distribution
χ^2_θ	$[-]$ chi-squared statistic of parameter estimate
∇	$[-]$ gradient operator in parameter space

Greek symbols – vectors and matrices

$\boldsymbol{\theta}$	$[N_\theta \times 1]$ column vector of parameters in parameter space Θ
$\boldsymbol{\theta}^*$	$[N_\theta \times 1]$ column vector of target parameters in parameter space Θ
$\hat{\boldsymbol{\theta}}$	$[N_\theta \times 1]$ maximum likelihood estimate for $\boldsymbol{\theta} \in \Theta$
$\boldsymbol{\Lambda}$	$[N_\theta \times N_\theta]$ diagonal matrix whose ii-th element is λ_i
$\boldsymbol{\Sigma}$	$[N_y \times N_y]$ covariance of measurement error for sample $\mathbf{y}$
$\boldsymbol{\varphi}$	$[N_u + 1 \times 1]$ experiment design vector
$\boldsymbol{\varphi}^*$	$[N_u + 1 \times 1]$ D-optimal experiment design vector
$\boldsymbol{\omega}$	$[N_\theta \times 1]$ column vector of parameters in parameter space Ω
$\hat{\boldsymbol{\omega}}$	$[N_\theta \times 1]$ maximum likelihood estimate for $\boldsymbol{\omega} \in \Omega$

Sub- and Superscript

BA	Benzoic Acid
EB	Ethyl Benzoate
Et	Ethanol
W	Water

Abbreviations

MBDoE	Model-Based Design of Experiments
RP	Reparametrization

8. References

[1] Echtermeyer, A.; Amar, Y.; Zakrzewski, J.; Lapkin, A. Self-Optimisation and Model-Based Design of Experiments for Developing a C-H Activation Flow Process. *Beilstein J. Org. Chem.* **2017**, *13*, 150–163. DOI: 10.3762/bjoc.13.18

[2] McMullen, J. P.; Jensen, K. F. Rapid Determination of Reaction Kinetics with an Automated Microfluidic System. *Org. Process Res. Dev.* **2011**, *15* (2), 398–407. DOI: 10.1021/op100300p

[3] Bournazou, M. N. C.; Barz, T.; Nickel, D. B.; Cárdenas, D. C. L.; Glauche, F.; Knepper, A.; Neubauer, P. Online Optimal Experimental Re-Design in Robotic Parallel Fed-Batch Cultivation Facilities. *Biotechnol. Bioeng.* **2016**, *114* (3), 610–619. DOI: 10.1002/bit.26192

[4] Franceschini, G.; Macchietto, S. Model-Based Design of Experiments for Parameter Precision: State of the Art. *Chem. Eng. Sci.* **2008**, *63* (19), 4846–4872. DOI: 10.1016/j.ces.2007.11.034

[5] Bard, Y. *Nonlinear Parameter Estimation*; Academic Press, **1974**.

[6] Wilson, A. D.; Schultz, J. A.; Murphey, T. D. Trajectory Optimization for Well-Conditioned Parameter Estimation. *IEEE Trans. Autom. Sci. Eng.* **2015**, *12* (1), 28–36. DOI: 10.1109/TASE.2014.2323934

[7] White, A.; Tolman, M.; Thames, H. D.; Withers, H. R.; Mason, K. A.; Transtrum, M. K. The Limitations of Model-Based Experimental Design and Parameter Estimation in Sloppy Systems. *PLOS Comput. Biol.* **2016**, *12* (12). DOI: 10.1371/journal.pcbi.1005227

[8] Higham, N. J. *Accuracy and Stability of Numerical Algorithms*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, **1996**.

[9] López C., D. C.; Barz, T.; Körkel, S.; Wozny, G. Nonlinear Ill-Posed Problem Analysis in Model-Based Parameter Estimation and Experimental Design. *Comput. Chem. Eng.* **2015**, *77*, 24–42. DOI: 10.1016/j.compchemeng.2015.03.002

[10] Barz, T.; López C., D. C.; Cruz Bournazou, M. N.; Körkel, S.; Walter, S. F. Real-Time Adaptive Input Design for the Determination of Competitive Adsorption Isotherms in Liquid Chromatography. *Comput. Chem. Eng.* **2016**, *94*, 104–116. DOI: 10.1016/j.compchemeng.2016.07.009

[11] Barz, T.; Cárdenas, D. C. L.; Arellano-Garcia, H.; Wozny, G. Experimental Evaluation of an Approach to Online Redesign of Experiments for Parameter Determination. *AIChE J.* **2013**, *59* (6), 1981–1995. DOI:10.1002/aic.13957

[12] Hansen, P. C. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*; SIAM, **2005**.

[13] Bardow, A. Optimal Experimental Design of Ill-Posed Problems: The METER Approach. *Comput. Chem. Eng.* **2008**, *32* (1), 115–124. DOI: 10.1016/j.compchemeng.2007.05.004

[14] Agarwal, A. K.; Brisk, M. L. Sequential Experimental Design for Precise Parameter Estimation. 1. Use of Reparameterization. *Ind. Eng. Chem. Process Des. Dev.* **1985**, *24* (1), 203–207. DOI: 10.1021/i200028a034

[15] Espie, D. M.; Macchietto, S. Nonlinear Transformations for Parameter Estimation. *Ind. Eng. Chem. Res.* **1988**, *27* (11), 2175–2179. DOI: 10.1021/ie00083a037

[16] Asprey, S. P.; Naka, Y. Mathematical Problems in Fitting Kinetic Models - Some New Perspectives. *J. Chem. Eng. Jpn.* **1999**, *32* (3), 328–337. DOI: 10.1252/jcej.32.328

[17] Schwaab, M.; Pinto, J. C. Optimum Reference Temperature for Reparameterization of the Arrhenius Equation. Part 1: Problems Involving One Kinetic Constant. *Chem. Eng. Sci.* **2007**, *62* (10), 2750–2764. DOI: 10.1016/j.ces.2007.02.020

[18] Schwaab, M.; Lemos, L. P.; Pinto, J. C. Optimum Reference Temperature for Reparameterization of the Arrhenius Equation. Part 2: Problems Involving Multiple Reparameterizations. *Chem. Eng. Sci.* **2008**, *63* (11), 2895–2906. DOI: 10.1016/j.ces.2008.03.010

[19] Buzzi-Ferraris, G.; Manenti, F. Kinetic Models Analysis. *Chem. Eng. Sci.* **2009**, *64* (5), 1061–1074. DOI: 10.1016/j.ces.2008.10.062

[20] 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* (15), 1923–1929. DOI: 10.1093/bioinformatics/btp358

[21] Pukelsheim, F. *Optimal Design of Experiments*; Society for Industrial and Applied Mathematics: New York, United States of America, **2006**.

[22] Saltelli, A.; Chan, K.; Scott, E. M. *Sensitivity Analysis*; Wiley: Chichester; New York, **2000**.

[23] Galvanin, F.; Macchietto, S.; Bezzo, F. Model-Based Design of Parallel Experiments. *Ind. Eng. Chem. Res.* **2007**, *46* (3). DOI: 10.1021/ie0611406

[24] Rimensberger, T.; Rippin, D. W. T. "Sequential Experimental Design for Precise Parameter Estimation. 1. Use of Reparameterization". Comments. *Ind. Eng. Chem. Process Des. Dev.* **1986**, *25* (4), 1042–1044. DOI: 10.1021/i200035a034

[25] Fedorov, V. V. *Theory Of Optimal Experiments*, 1972nd ed.; Academic Press, **1972**.

[26] Pipus, G.; Plazl, I.; Koloini, T. Esterification of Benzoic Acid in Microwave Tubular Flow Reactor. *Chem. Eng. J.* **2000**, *76* (3), 239–245. DOI: 10.1016/S1385-8947(99)00171-0

9. Tables with headings

Tab. 1. Non-RP campaign. Parameter estimates, 95% confidence intervals and correlation coefficient in the course of the simulated campaign. Parameter estimation and MBDoE problems are solved in the original parameter space Θ. The condition number of the log-likelihood function in Θ is reported for each iteration of the algorithm (i.e. after the collection of every sample).

Online RP Inactive

Samples collected	Estimates $\widehat{\boldsymbol{\theta}} = [\theta_1, \theta_2]$ with 95% confidence intervals	Correlation coefficient c_{12}	p-value of target parameters $\boldsymbol{\theta}^*$	Condition number κ in Θ
1	[- , -]	-	-	-
2	[14.15 ± 2.11 , 7.23 ± 1.41]	0.9998	0.00%	$1.2{\cdot}10^4$
3	[13.72 ± 1.32 , 7.08 ± 0.85]	0.9997	0.00%	$8.1{\cdot}10^3$
4	[14.44 ± 0.99 , 7.32 ± 0.67]	0.9998	0.00%	$1.1{\cdot}10^4$
5	[14.28 ± 0.78 , 7.26 ± 0.52]	0.9997	0.00%	$7.2{\cdot}10^3$
6	[14.38 ± 0.73 , 7.29 ± 0.49]	0.9998	0.00%	$9.5{\cdot}10^3$
7	[14.85 ± 0.65 , 7.45 ± 0.43]	0.9997	0.00%	$8.2{\cdot}10^3$
8	[15.05 ± 0.61 , 7.52 ± 0.41]	0.9998	0.00%	$9.5{\cdot}10^3$

Tab. 2. RP campaign. Parameter estimates, 95% confidence intervals and correlation coefficient in the course of the campaign. Parameter estimation and MBDoE problems are solved in the transformed parameter space Ω. The condition number of the log-likelihood function in Ω is reported in the table at each iteration of the algorithm (i.e. after the collection of every sample). The condition number is computed at the *parameter estimation step* (i.e. after the *reparametrization step* has occurred)

Online RP Active

Samples collected	Estimates $\widehat{\boldsymbol{\theta}} = [\theta_1, \theta_2]$ with 95% confidence intervals	Correlation coefficient c_{12}	p-value of target parameters $\boldsymbol{\theta}^*$	Condition number κ in Ω
1	[- , -]	-	-	-
2	[11.57 ± 3.69 , 6.33 ± 1.27]	0.9998	13.16%	$7.9{\cdot}10^5$
3	[14.00 ± 1.35 , 7.16 ± 0.43]	0.9985	0.08%	$8.8{\cdot}10^5$
4	[13.84 ± 1.74 , 7.11 ± 0.59]	0.9997	17.30%	$4.5{\cdot}10^4$
5	[14.96 ± 1.35 , 7.49 ± 0.46]	0.9996	76.19%	$1.6{\cdot}10^1$
6	[15.00 ± 1.28 , 7.50 ± 0.43]	0.9997	69.35%	$1.0{\cdot}10^0$
7	[15.62 ± 1.13 , 7.72 ± 0.38]	0.9997	81.78%	$1.0{\cdot}10^0$
8	[15.62 ± 1.09 , 7.72 ± 0.37]	0.9997	78.13%	$1.0{\cdot}10^0$

10.Figure legends

Fig. 1. Proposed framework for the online identification of models in automated model identification platforms. After each sample, the parametrization matrix $\mathbf{G}$ is updated with the aim of minimizing the condition number of the Hessian associated to the parameter estimation problem. The online update of the parametrization is performed to reduce the risk of numerical failures at the parameter estimation and optimal MBDoE steps in the procedure.

Fig. 2. Condition number κ after each collected sample in the non-RP campaign (dotted line) and in the RP campaign (solid line). The condition number for the RP-campaign is computed at the *parameter estimation step* of the algorithm, i.e. after the update of the transformation matrix $\mathbf{G}$.

Fig. 3. Estimates of parameters θ_1 and θ_2 in the non-RP campaign (dotted line) and in the RP campaign (solid line) at each collected sample in the campaigns: a) estimated values; b) $\pm 95\%$ confidence intervals.

Fig. 4. Final parameter estimates and related 95% confidence ellipsoids in non-RP campaign (dotted ellipsoid) and RP campaign (solid ellipsoid). The target parameter value is highlighted by a star-shaped symbol.

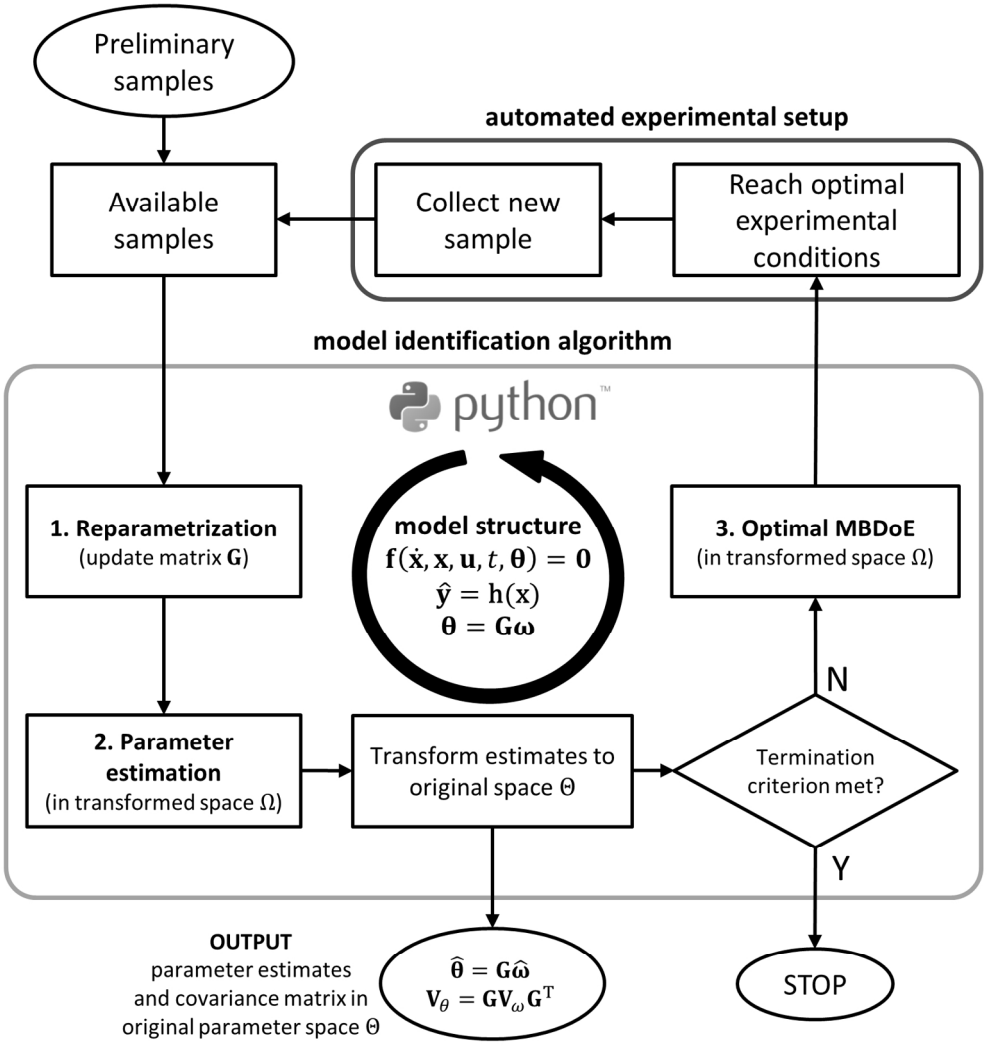

Fig. 1. Proposed framework for the online identification of models in automated model identification platforms. After each sample, the parametrization matrix **G** is updated with the aim of minimizing the condition number of the Hessian associated to the parameter estimation problem. The online update of the parametrization is performed to reduce the risk of numerical failures at the parameter estimation and optimal MBDoE steps in the procedure.

149x160mm (300 x 300 DPI)

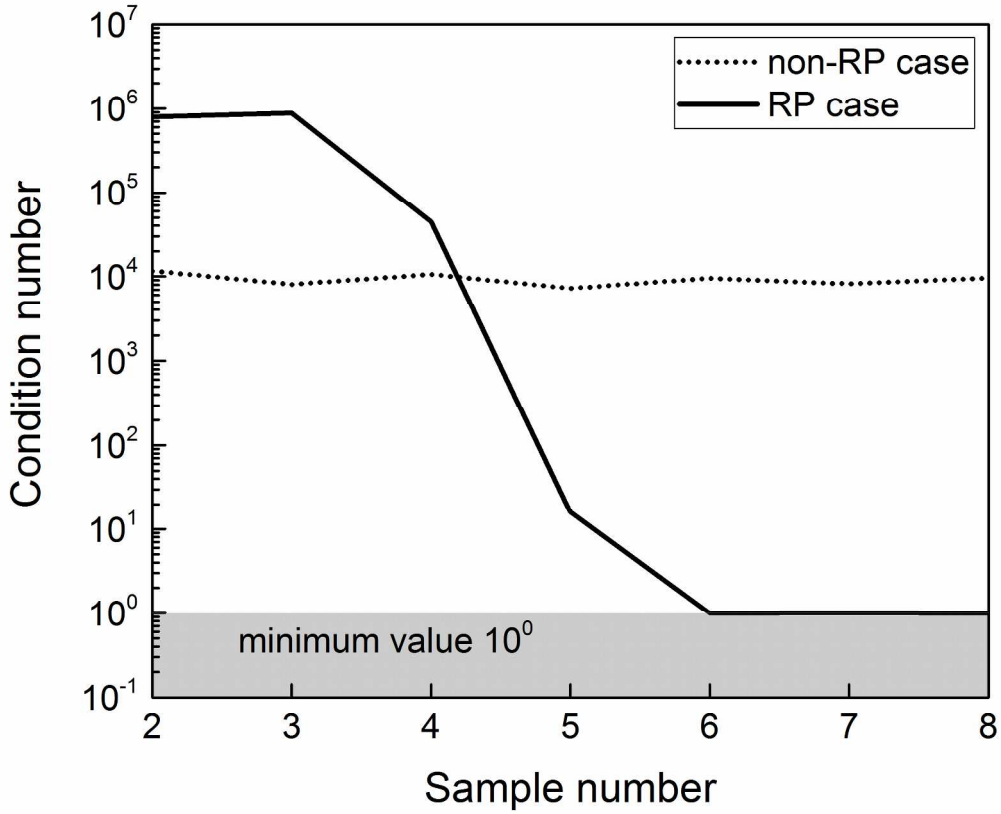

Fig. 2. Condition number κ after each collected sample in the non-RP campaign (dotted line) and in the RP campaign (solid line). The condition number for the RP-campaign is computed at the parameter estimation step of the algorithm, i.e. after the update of the transformation matrix **G**.

222x181mm (300 x 300 DPI)

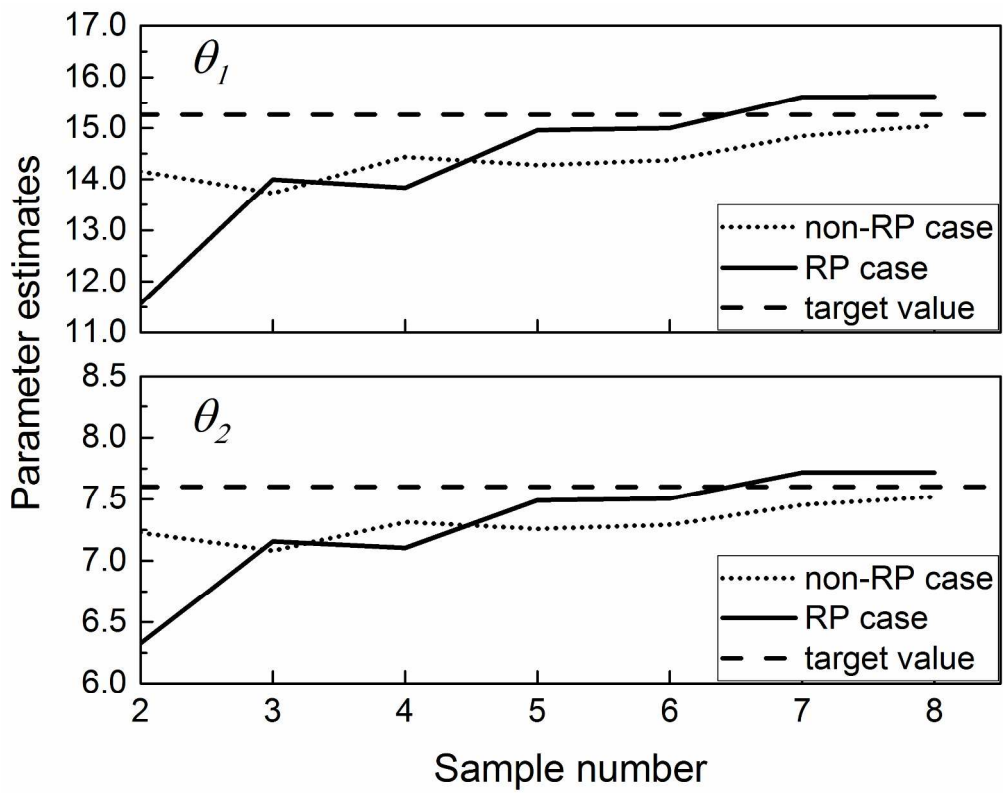

Fig. 3. Estimates of parameters θ_1 and θ_2 in the non-RP campaign (dotted line) and in the RP campaign (solid line) at each collected sample in the campaigns: a) estimated values; b) ±95% confidence intervals.

237x187mm (300 x 300 DPI)

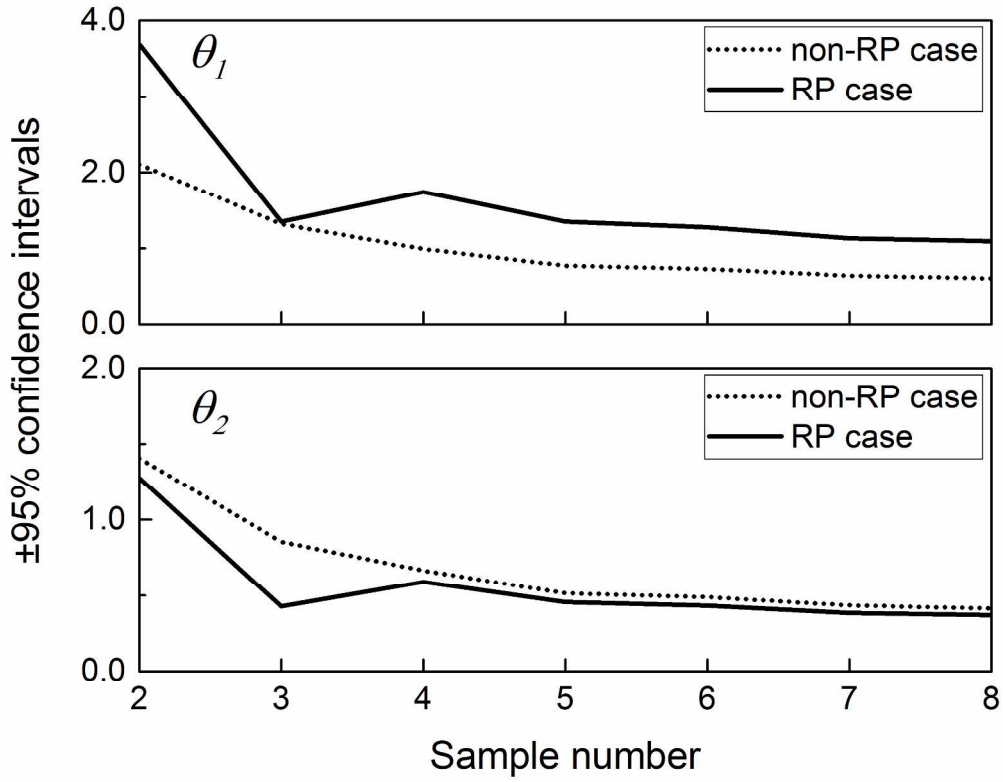

Fig. 3. Estimates of parameters θ_1 and θ_2 in the non-RP campaign (dotted line) and in the RP campaign (solid line) at each collected sample in the campaigns: a) estimated values; b) ±95% confidence intervals.

240x187mm (300 x 300 DPI)

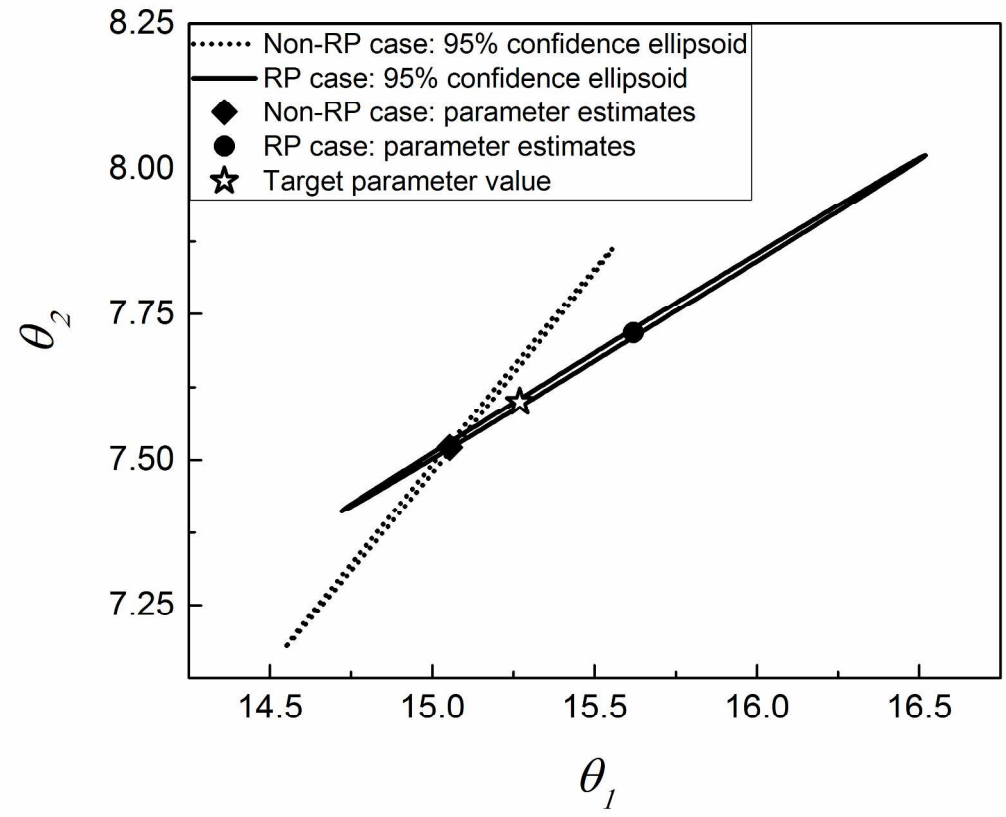

Fig. 4. Final parameter estimates and related 95% confidence ellipsoids in non-RP campaign (dotted ellipsoid) and RP campaign (solid ellipsoid). The target parameter value is highlighted by a star-shaped symbol.

229x186mm (300 x 300 DPI)