Integration based profile likelihood calculation for PDE constrained parameter estimation problems

R Boiger[1], J Hasenauer[2,3], S Hroß[2,3] and B Kaltenbacher[1]

[1] Alpen-Adria-Universität Klagenfurt, Austria
[2] Helmholtz Zentrum München, Institute of Computational Biology, Germany
[3] Technische Universität München, Department of Mathematics, Germany

Abstract

Partial differential equation (PDE) models are widely used in engineering and natural sciences to describe spatio-temporal processes. The parameters of the considered processes are often unknown and have to be estimated from experimental data. Due to partial observations and measurement noise, these parameter estimates are subject to uncertainty. This uncertainty can be assessed using profile likelihoods, a reliable but computationally intensive approach.

In this paper, we present the integration based approach for the profile likelihood calculation developed by Chen and Jennrich [5] and adapt it to inverse problems with PDE constraints. While exiting methods for profile likelihood calculation in parameter estimation problems with PDE constraints rely on repeated optimization, the proposed approach exploits a dynamical system evolving along the likelihood profile. We derive the dynamical system for the unreduced estimation problem, prove convergence and study the properties of the integration based approach for the PDE case. To evaluate the proposed method, we compare it with state-of-the-art algorithms for a simple reaction-diffusion model for a cellular patterning process. We observe a good accuracy of the method as well as a significant speed up as compared to established methods. Integration based profile calculation facilitates rigorous uncertainty analysis for computationally demanding parameter estimation problems with PDE constraints.

1 Introduction

Engineering, physics, biology and adjacent fields employ PDE models for the mathematical description of involved processes. These models often contain unknown parameters which have to be inferred from experimental data. The corresponding parameter estimation problems are potentially ill-posed due to limited and noise-corrupted experimental data [12]. Due to the ill-posedness a comprehensive uncertainty analysis is crucial. In particular, we refer to [8] for an overview on inverse problems in systems biology with an emphasis on regularization aspects. Since we here deal with finite dimensional parameter spaces, regularization (see, e.g., [7]) is not required. Still we face the difficulty that some parameters might not be uniquely determined from the given noisy measurements and due to the strong nonlinearity of the problem, this indeterminacy might not be detectable by just considering the nullspace of the linearized forward operator. Thus we here rely on the concept of practical identifiability and profile likelihoods that fully account for nonlinearity. So far, the literature on profile likelihoods appears to mainly concentrate on (finite dimensional) statistics, as well as applications in systems biology, geography and econometrics. To the best of our knowledge, their use in inverse problems involving models in infinite dimensional spaces, especially to parameter identification problems in PDEs, has not been investigated yet.

In a statistical framework, parameter and prediction uncertainties can be quantified in terms of confidence and credible intervals. Confidence and credible intervals capture the range of plausible parameter and model predictions in accordance with a predefined statistical measure, e.g., the likelihood ratio. For the construction of confidence and credible intervals, local approximations [25, 26], bootstrapping [19], Bayesian methods [33] and profile likelihoods [26] are employed. Local approximation such as the Wald approximation [25] and the Fisher Information Matrix (FIM) based approximation [26] are computationally efficient but merely provide rough estimates of confidence intervals. Bootstrapping provides non-local estimates but should only be applied to models without practical non-identifiablities [9]. Bayesian methods and profile likelihoods appear to be most reliable and consistent [17, 18, 28].

Bayesian methods construct representative samples from the posterior distribution, thereby assessing the uncertainty of all parameters and model predictions simultaneously [20]. Profile likelihood methods explore the uncertainty of individual parameters and model predictions using repeated local optimizations. The credible intervals computed using Bayesian methods employ marginalization, while confidence intervals computed using profile likelihoods rely on maximum projections. For well-posed problems, it follows from asymptotic normality that profile likelihoods and marginals are identical up to a scaling constant. Even for finite sample sizes, the agreement of profile likelihoods and marginals is usually rather high (see, e.g.,[18, 28, 17]). Raue et al. [28] demonstrated that profile likelihood based confidence intervals can be advantageous as the coverage of regions with high likelihood values is ensured. In addition, the calculation of profile likelihoods tends to be computationally more tractable than the sampling of the posterior distribution [17, 18, 28]. This also holds if sophisticated sampling procedures [10, 11, 29] are used. Nevertheless, for computationally demanding problems, also the application of classical profile likelihood methods is prohibitive [16, 17, 24].

To improve the computational efficiency of profile likelihood calculations, Chen and Jennrich [5] proposed an integration based approach. This approach relies on a differential algebraic equation (DAE) which evolves along the profile likelihood. The trajectories of this systems provide the parameter profile without the need for repeated optimization. Mass matrix and vector field of the DAE are computed from the gradient and hessian of the objective function. Chen and Jennrich [5] obtained promising results for simple likelihood functions. In the last years also the application to ordinary differential equation (ODE) models has been discussed [23].

In this paper, we will generalize integration based profile likelihood calculation to PDE constrained parameter estimation problems. We will introduce a reduced and a full formulation for a statistically motivated objective function and discuss their properties. As the calculation of the Hessian is potentially computationally intensive, approximation will be considered and combined with a retraction term. The different approaches will be illustrated and evaluated using an example from systems biology.

The remainder of this paper is organized as follows: In Section 2 we will introduce the considered class of mathematical models and observation operator. The parameter estimation problem and uncertainty analysis using profile likelihoods will be outlined in Section 3. In Section 4 and 5 the integration based profile likelihood calculation for the reduced and the full problem are presented. The relation of these two approaches is discussed in Section 6. The proposed integration based profile likelihood calculation for PDE models is evaluated in Section 7 for a model of gradient formation in fission yeast. The paper concludes with a discussion of the results and an outlook in Section 8.

2 Mathematical model

We consider parameter estimation in partial differential equation models

$$\begin{aligned} &u_t + C(\theta, u) = f(\theta) \text{ in }]0, T[\\ &u(0) = u_0, \end{aligned} \tag{1}$$

with state variable $u \in V$, defined over a spatial domain, and parameter vector $\theta \in \Theta \subseteq \mathbb{R}^n$. The operator $C(\theta, .) : V \to V^*$, mapping from a separable, reflexive Banach space V into its dual V^*, is

equipped with appropriate boundary conditions, where $V \subset H \cong H^* \subset V^*$ is a Gelfand triple such that V is imbedded continuously and densely into a Hilbert space H. To guarantee the existence of a weak solution $u \in W(0,T) = L^2(0,T;V) \cap H^1(0,T;V^*)$ of (1), according to ([34], p. 770 ff.), we assume that the operator C meets the following assumption:

Assumption 2.1 (Existence of a weak solution)**.**

- $u_0 \in H$ *and* $f(\theta) \in L^2([0,T];V^*)$ *are given.*
- $C(\theta,.)$ *is monotone and hemicontinuous.*
- $C(\theta,.)$ *is coercive, i.e. there exist* c_0 *and* c_1 *such that* $\langle C(\theta,u),u\rangle_{V^*,V} \geq c_0\|u\|_V^2 - c_1$.
- $C(\theta,.)$ *satisfies the growth condition, i.e. there exists a nonnegative function* $c_2 \in L^2(0,T)$ *and a constant* $c_3 > 0$*, such that* $\|C(\theta,u)(t)\|_{V^*} \leq c_2(t) + c_3\|u(t)\|_V$ *for all* $u \in V$ *and* $t \in]0,T[$.
- *The function* $t \mapsto \langle C(\theta,u)(t),v\rangle_{V^*,V}$ *is measurable on* $]0,T[$ *for all* $u,v \in V$.

Assumption 2.1 holds for models from a broad range of applications [32] and ensures the existence of a parameter-to-state map

$$S : \mathbb{R}^{n_\theta} \to W(0,T), \text{ with } u = S(\theta) \text{ solving (1).} \tag{2}$$

As the measurement of u can be limited by experimental technologies, we consider potentially partial observations,

$$y = Q(\theta,u). \tag{3}$$

The observation operator $Q(\theta,.) : W(0,T) \to \mathbb{R}^K$ maps u onto the observation $y \in \mathbb{R}^K$. The observation y is a collection of different scalar observables measured at different time points. The index k enumerates all the combinations of observables and time points, $y_k = (Q(\theta,u))_k = Q_k(\theta,u)$ for $k = 1,\ldots,K$.

In practice the observations are corrupted by measurement noise. The noise corrupted measurement of the observable y_k is denoted by $\overline{y}_k$. For additive, normally distributed measurement noise it holds that

$$\overline{y}_k = y_k + \varepsilon_k \text{ with } \varepsilon_k \sim \mathcal{N}(0,\sigma_k^2(\theta)). \tag{4}$$

The parameters of the noise model, here the variance $\sigma_k^2(\theta)$, are potentially unknown and can depend on the parameter.

3 Parameter estimation problem and uncertainty analysis

We estimate the parameters using a likelihood-based approach. The likelihood is the conditional probability of observing the measured data $\overline{y}_k$, $k = 1,\ldots,K$, given the parameter vector θ. The likelihood of observing the measured data depends implicitly on the noise model and is e.g. for additive, normally distributed measurement noise given by

$$\mathrm{L}(\theta) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}\sigma_k(\theta)} \exp\left(-\frac{1}{2}\left(\frac{\overline{y}_k - Q_k(\theta,S(\theta))}{\sigma_k(\theta)}\right)^2\right).$$

Remark 3.1. *The methods and results we will present do not assume a particular noise model or likelihood function. We merely assume that the likelihood function is twice continuously differentiable.*

The parameter vector $\hat{\theta}$ which maximizes $\mathrm{L}(\theta)$ is the maximum likelihood estimate. To improve the numerical evaluation and the optimizer convergence, the maximum likelihood estimate is usually determined by minimizing the negative log-likelihood function,

$$\mathbf{j}(\theta) = -\log \mathrm{L}(\theta) = \frac{1}{2}\sum_{k=1}^{K}\left(\log\left(2\pi\sigma_k^2(\theta)\right) + \left(\frac{\overline{y}_k - Q_k(\theta,S(\theta))}{\sigma_k(\theta)}\right)^2\right), \tag{5}$$

with its non-reduced counterpart

$$\mathbf{J}(\theta, u) = \frac{1}{2}\sum_{k=1}^{K}\left(\log\left(2\pi\sigma_k^2(\theta)\right) + \left(\frac{\bar{y}_k - Q_k(\theta, u)}{\sigma_k(\theta)}\right)^2\right).$$

This yields the PDE constrained optimization problem

$$\begin{aligned}\min_{\theta\in\Theta, u\in W(0,T)} \mathbf{J}(\theta, u)\\ \text{s.t. } u_t + C(\theta, u) = f(\theta) \text{ in }]0, T[\\ u(0) = u_0.\end{aligned} \tag{6}$$

The maximum likelihood estimate $\hat{\theta}$, i.e. the optimum of (6), is potentially non-unique and might strongly depend on the measurement noise. To assess the parameters and prediction uncertainties we consider confidence regions and confidence intervals.

Definition 1 (Confidence region)**.**
For the parameter vector $\theta \in \Theta$ we define the confidence region to the confidence level α as

$$\begin{aligned}\mathrm{CR}_\alpha &= \left\{\theta \in \Theta \,\middle|\, \frac{\mathrm{L}(\theta)}{\mathrm{L}(\hat{\theta})} \geq \exp\left(-\frac{\Delta_\alpha}{2}\right)\right\},\\ &= \left\{\theta \in \Theta \,\middle|\, 2\left(\mathbf{j}(\theta) - \mathbf{j}(\hat{\theta})\right) \leq \Delta_\alpha\right\},\end{aligned} \tag{7}$$

with Δ_α denoting the αth-percentile of the χ^2 distribution with one degree of freedom.

From the confidence regions, the confidence intervals for individual model properties $\mathbf{G}(\theta, u)$, with $\mathbf{G} : \Theta \times V \mapsto \mathbb{R}$, can be derived. The reduced form of $\mathbf{G}(\theta, u)$ is denoted by $\mathbf{g}(\theta) = \mathbf{G}(\theta, S(\theta))$, with $\mathbf{g} : \Theta \mapsto \mathbb{R}$. Model properties are for instance individual parameters, functions of parameters or properties of the solution of the model.

Definition 2 (Confidence interval)**.**
The confidence interval for a model property is the projection of CR_α *onto the range* $\mathbf{g}(\Theta)$ *of* $\mathbf{g}$*.*

$$\mathrm{CI}_{\alpha,\mathbf{g}} = P_{\mathbf{g}(\Theta)}\mathrm{CR}_\alpha = \{c \,|\exists\theta \in \mathrm{CR}_\alpha \,:\, \mathbf{g}(\theta) = c\} \tag{8}$$

The evaluation of the confidence region requires the calculation of level sets of likelihood functions. For problems with $n_\theta \gg 1$ this is non-trivial. To determine confidence intervals without calculating confidence regions, profile likelihoods [26] can be used. The profile likelihood for a scalar function $\mathbf{g}$, $\mathrm{PL}_\mathbf{g}(c)$, is the maximal likelihood value for $\mathbf{g}(\theta) = c$ [4].

Definition 3 (Profile likelihood)**.**
For the scalar function $\mathbf{g}$ *we define the profile likelihood as*

$$\mathrm{PL}_\mathbf{g}(c) = \max_{\theta\in\Theta} \mathrm{L}(\theta) \text{ subject to } \mathbf{g}(\theta) = c. \tag{9}$$

For values c outside the range of $\mathbf{g}$*, we set* $\mathrm{PL}_\mathbf{g}(c) = 0$*.*

In other words the profile likelihood provides the maximum projection of the likelihood along $\mathbf{g}(\Theta)$. Accordingly, the confidence interval for $\mathbf{g}$ follows from (8) as

$$\mathrm{CI}_{\alpha,\mathbf{g}} = \left\{c \,\middle|\, \frac{\mathrm{PL}_\mathbf{g}(c)}{\mathrm{L}(\hat{\theta})} \geq \exp\left(-\frac{\Delta_\alpha}{2}\right)\right\}.$$

The relation among likelihood function, confidence region/intervals and profile likelihoods is illustrated in Figure 1. Note that the confidence intervals for the individual parameters are obtained by the projection of the confidence region as well as by thresholding the profile likelihood. For $\mathbf{g}(\theta) = \theta_j$ this provides the confidence interval of parameter θ_j, while for other choices of $\mathbf{g}$ more involved parameter dependent model properties can be assessed, e.g. the product of two parameters.

The confidence intervals to a confidence level α can be bounded or unbounded:

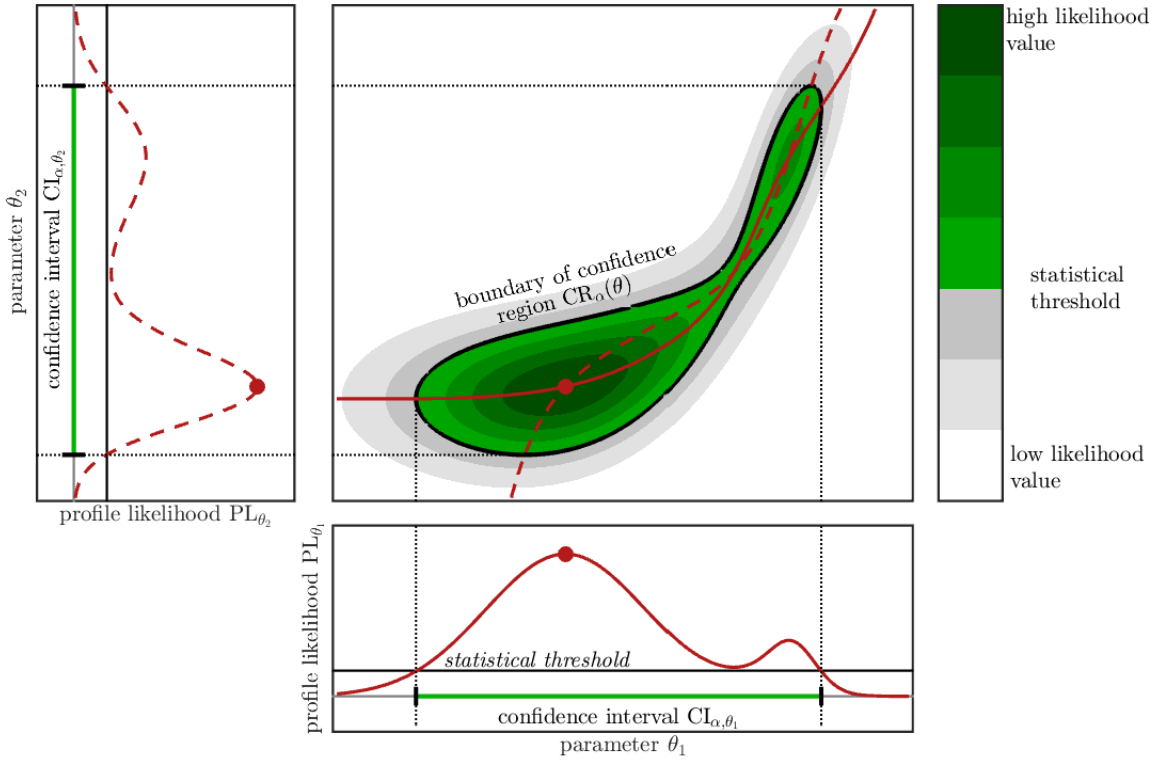

Figure 1: **Illustration of confidence region, confidence intervals, profile likelihoods and their relation.** (big panel) Likelihood function landscape (shading), confidence region (■) and profile likelihood path θ_c (θ_1:━;θ_2:╍). (small panels) Profile likelihood ratio (━) and confidence interval (━) for θ_1 and θ_2. The relation of different quantities is indicated using dotted lines. The significance threshold is indicated in all three figures as solid black line.

Definition 4 (Practical identifiability)**.**
A model property $\mathbf{g}$ *is called practically identifiable if its confidence interval* $\mathrm{CI}_{\alpha,\mathbf{g}}$ *is bounded; otherwise it is called practically non-identifiable.*

For a detailed introduction to profile likelihoods and confidence intervals we refer to the statistical literature (e.g., [25, 26]) and the applications (e.g. [16, 27]).

4 Profile likelihood calculation for the reduced problem

In this section, we introduce optimization and integration based profile likelihood calculation for the reduced form of PDE constrained optimization problems. The formulation of the integration based profile likelihood calculation method is adapted from the results of Chen and Jennrich [5]. In particular we establish its validity for function spaces. To keep the exposition focused on the profile likelihood computations, we consider the case without constraints on the parameters, i.e. $\theta \in \Theta = \mathbb{R}^n$. Constraints on the parameters would lead to additional Lagrange multipliers in the first order optimality conditions and require regularity assumptions (constraint qualifications) to justify existence of these multipliers.

Since our analysis will rely on differentiation of the first order necessary optimality conditions, we will make the following assumptions on smoothness of the involved functions

Assumption 4.1. $C\colon \mathbb{R}^{n_\varphi} \times V \to V^*$, $f\colon \mathbb{R}^{n_\varphi} \to V^*$ *and* $J\colon \mathbb{R}^{n_\theta} \times W(0,T) \to \mathbb{R}$ *are twice continuously differentiable.*

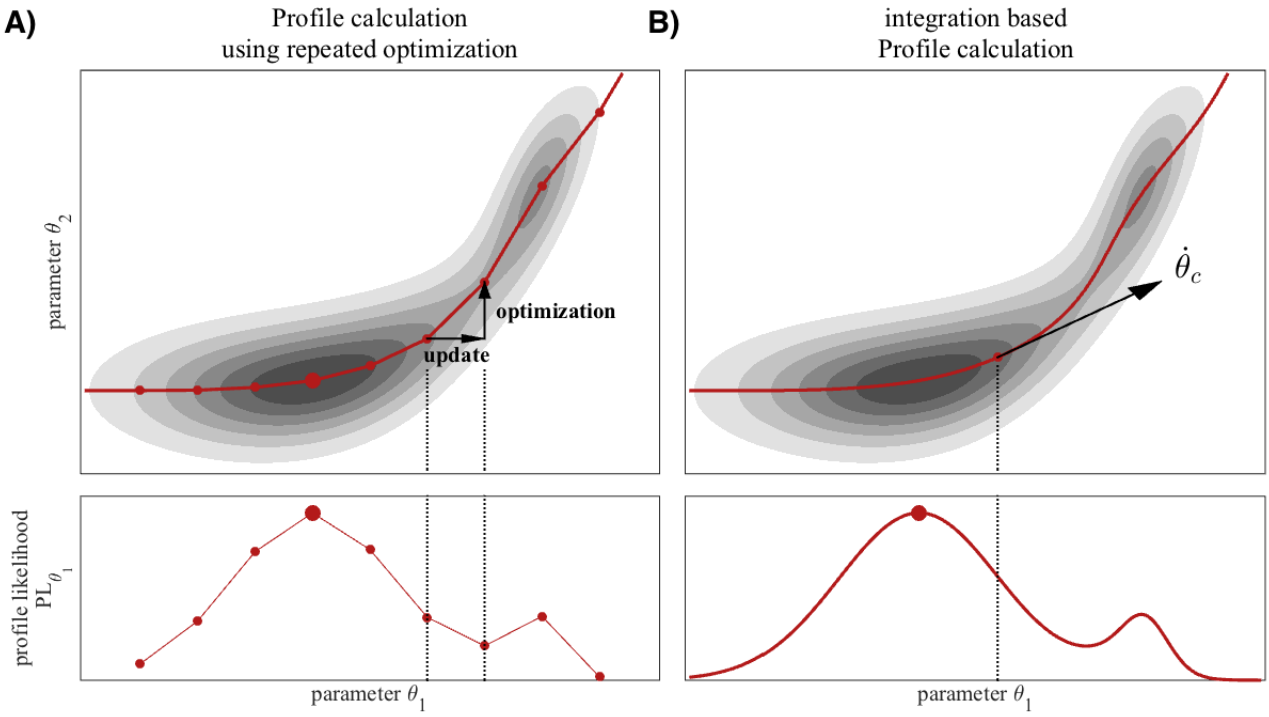

Figure 2: **Illustration of optimization based profile likelihood calculation (upper panels) and integration based profile likelihood calculation (lower panels).** (upper panel in A) Optimization based profile likelihood calculation for likelihood function (shading) using update and re-optimization step (arrows). (upper panel in B) Evaluation points (•) and approximation of the profile likelihood (━). (lower panel in A) Integration based profile likelihood calculation for likelihood function (shading) using continuous system with derivative (arrow) tangential to the parameter trajectory (━). (lower panel in B) Profile likelihood (━) obtained using integration based method.

4.1 Optimization based profile likelihood calculation

Optimization based methods approximate the profile likelihood $\mathrm{PL}_{\mathbf{g}}(c)$ by evaluating it on a grid $\{c_l\}_l$ (Figure 2 upper panels). For each point c the reduced negative log-likelihood function is minimized,

$$\begin{aligned} &\min_{\theta\in\Theta} \mathbf{j}(\theta) \\ &\text{s.t. } \mathbf{g}(\theta) = c. \end{aligned} \tag{10}$$

This minimization yields the optimal parameter vector, $\theta_c := \hat{\theta}(c)$, and the corresponding value of the negative log-likelihood function, $\mathbf{j}(\theta_c)$. It holds that $\mathrm{PL}_{\mathbf{g}}(c) = \exp(-\mathbf{j}(\theta_c))$.

State-of-the-art methods construct the grid $\{c_l\}_l$ iteratively, starting at the optimal parameter vector $\hat{\theta}$ of (6) with $\hat{c} = \mathbf{g}(\hat{\theta})$ [27]. An iteration consists of two steps: (i) the update of the constraint c_l using an adaptive approach; and (ii) the local optimization of the parameters. The adaptation controls the change in the objective function, $\mathbf{j}(\theta_{c_{l-1}})$ to $\mathbf{j}(\theta_{c_l})$, and the number of iterations. The starting point $\theta_{c_l}^{(0)}$ of the local optimization for c_l is constructed from previous points. Most implementations use as starting point

1. **0th order proposal:** the optimal point for c_{l-1}, $\theta_{c_l}^{(0)} = \theta_{c_{l-1}}$, or
2. **1st order proposal:** the linear extrapolation based on the optimal points for c_{l-1} and c_{l-2},

$$\theta_{c_l}^{(0)} = \theta_{c_{l-1}} + \frac{c_l - c_{l-1}}{c_{l-1} - c_{l-2}}(\theta_{c_{l-1}} - \theta_{c_{l-2}}).$$

The 0th order proposal is illustrated in Figure 2 A (upper panel). In practice the 1st order proposal, which uses additional topological information, yields starting points which are closer to the optimum θ_c. Accordingly, this approach tends to be computationally more efficient.

Optimization-based profile likelihood calculation is computationally efficient compared to other uncertainty analysis methods [28, 18, 17]. It, however, becomes computationally demanding if the number of necessary iterations increases or an individual local optimization is computationally expensive. The number of iterations is influenced by the structure of the objective function landscape, e.g., non-identifiabilities. The computational complexity of the local optimization is determined by the computation time of the forward problems (and its derivatives). Both are issues for a range of practical applications including PDE constrained problems [16].

4.2 Integration based profile likelihood calculation

Integration based profile likelihood calculation addresses the drawbacks of optimization based methods by exploiting the differential geometry of the reduced optimization problem [5]. This is achieved by considering the Lagrange function for (10),

$$\ell(\theta, \lambda) = \mathbf{j}(\theta) + \lambda(\mathbf{g}(\theta) - c),$$

in which $\lambda \in \mathbb{R}$ denotes the Lagrange multiplier. From the Lagrange function the first order optimality conditions,

$$\begin{aligned} \nabla_\theta \mathbf{j}(\theta) + \lambda \nabla_\theta \mathbf{g}(\theta) &= 0 \\ \mathbf{g}(\theta) &= c, \end{aligned} \tag{11}$$

can be derived. This system of equations describes the dependence of the minimizing parameter vector θ and the Lagrange multiplier λ on c. Therefore we use the notation $\theta_c := \theta(c)$ and $\lambda_c := \lambda(c)$. The differentiation of (11) with respect to c yields an evolution equation for the pair (θ_c, λ_c),

$$\underbrace{\begin{pmatrix} \nabla_\theta^2 \mathbf{j}(\theta_c) + \lambda_c \nabla_\theta^2 \mathbf{g}(\theta_c) & \nabla_\theta \mathbf{g}(\theta_c) \\ \nabla_\theta \mathbf{g}(\theta_c)^T & 0 \end{pmatrix}}_{:= M_{\mathrm{red}}(\theta_c)} \begin{pmatrix} \dot{\theta}_c \\ \dot{\lambda}_c \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{12}$$

where $\dot{\theta}_c$ and $\dot{\lambda}_c$ are derivatives with respect to c. The solution of the differential algebraic equation (DAE) (12) for a starting point which solves (10) for $c = c_0$ yields the profile θ_c for $c \in [c_0, c_{\text{end}}]$.

Proposition 4.2. *Let $\mathbf{j} : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and let $(\theta_c, \lambda_c)_{c\in[c_0,c_1]}$ be a solution of (12) with initial data $(\theta_{c_0}, \lambda_{c_0})$ solving (11) for $c = c_0$.*

Then for all $c \in [c_0, c_1]$, (θ_c, λ_c) solves the optimality conditions (11).

Proof. For any fixed $c_1 > c_0$ we define $\Psi : [c_0, c_1] \to \mathbb{R}^{n+1}$ by $\Psi(c) = (\theta_c, \lambda_c)$ and $\Phi : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ by $\Phi(\theta, \lambda) = (\nabla_\theta \mathbf{j}(\theta) + \lambda \nabla_\theta \mathbf{g}(\theta), \mathbf{g}(\theta))^T$, so that we can rewrite (11) for $c \in [c_0, c_1]$ as

$$\Phi(\Psi(c)) - \begin{pmatrix} 0 \\ c \end{pmatrix} = 0 \quad \forall c \in [c_0, c_1]\,.$$

Under the differentiability assumptions made here this is equivalent to

$$\Phi(\Psi(c_0)) - \begin{pmatrix} 0 \\ c_0 \end{pmatrix} = 0 \;\text{ and }\; \frac{d\Phi}{d(\theta,\lambda)}(\Psi(c))\,\dot{\Psi}(c) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0 \quad \forall c \in [c_0, c_1]\,,$$

i.e., (12). □

The trajectory θ_c of (12) is the path in parameter space along which the minimum of the constrained optimization problem (10) is attained. The evaluation of the objective function $\mathbf{j}(\theta_c)$ along this trajectory yields the profile likelihood $\text{PL}_\mathbf{g}(c) = \exp(-\mathbf{j}(\theta_c))$. Accordingly, the profile likelihood can be computed without optimization. Instead, the update directions are determined by the derivatives of $\mathbf{g}$ and $\mathbf{j}$.

The numerical integration of (12) relies on the evaluation of the matrix-vector product $M_{\text{red}}(\theta_c)(\dot{\theta}_c, \dot{\lambda}_c)^T$. This matrix-vector product contains the terms $(\nabla^2_\theta \mathbf{j}(\theta_c) + \lambda_c \nabla^2_\theta \mathbf{g}(\theta_c))\dot{\theta}_c$, $\nabla_\theta \mathbf{g}(\theta_c)\dot{\lambda}_c$ and $\nabla_\theta g^T(\theta_c)\dot{\theta}_c$. As $\mathbf{g}$ and $\mathbf{j}$ are functions of the PDE solution, their derivatives depend on the parameter-to-state mapping $S(\theta)$. The sensitivities of the parameter-to-state map $S(\theta)$ can be calculated using forward sensitivity equations derived by differentiating (1) for $u = S(\theta_c)$. This differentiation yields the first and second order derivatives, $e = S_{\theta_i}(\theta)$ and $z = S_{\theta_i\theta_j}(\theta)$, for $i, j = 1, \ldots, n$, as solutions to,

$$\begin{cases} e_t + C_u(\theta, S(\theta))e = f_{\theta_i}(\theta) - C_{\theta_i}(\theta, S(\theta)) =: -h_i^\theta \\ e(0) = 0 \end{cases}$$

and

$$\begin{cases} z_t + C_u(\theta, S(\theta))z = & f_{\theta_i\theta_j}(\theta) - C_{\theta_i\theta_j}(\theta, S(\theta)) - C_{\theta_i u}(\theta, S(\theta))S_{\theta_j}(\theta) \\ & -C_{u\theta_j}(\theta, S(\theta))S_{\theta_i}(\theta) - C_{uu}(\theta, S(\theta))S_{\theta_i}(\theta)S_{\theta_j}(\theta) =: -h_{ij}^\theta \\ z(0) = 0. \end{cases} \tag{13}$$

In addition to solving the nonlinear PDE (1), this requires the solutions of n_θ linear PDEs for the first order sensitivities and $n_\theta(n_\theta + 1)/2$ linear PDEs for the second order sensitivities. An alternative to forward sensitivities is the evaluation of the aforementioned Hessian and gradient vector product by adjoint methods. These enable the computation of the objective function gradient by solving just one linearized PDE and computation of the Hessian of the objective function by solving two additional linearized PDEs. Namely, defining p as the solution of the adjoint equation

$$\begin{cases} p_t - C_u(\theta, S(\theta))^* p = \mathbf{J}_u(\theta, S(\theta)) \\ p(T) = 0 \end{cases} \tag{14}$$

(see also (23) below) and using the fact that $u = S(\theta)$ solves the PDE (1) followed by integration by parts, we obtain

$$\begin{aligned} \frac{\partial \mathbf{j}}{\partial \theta_i}(\theta) &= \frac{\partial}{\partial \theta_i}\Big(\mathbf{J}(\theta, S(\theta)) + \int_0^T \langle S(\theta)_t + C(\theta, S(\theta)) - f(\theta), p\rangle_{V^*,V} dt\Big) \\ &= \mathbf{J}_{\theta_i}(\theta, S(\theta)) + \mathbf{J}_u(\theta, S(\theta))S_{\theta_i}(\theta) + \int_0^T \Big(\langle S_{\theta_i}(\theta)_t + C_{\theta_i}(\theta, S(\theta)) + C_u(\theta, S(\theta))S_{\theta_i}(\theta) - f_{\theta_i}(\theta), p\rangle_{V^*,V}\Big) dt \\ &= \mathbf{J}_{\theta_i}(\theta, S(\theta)) + \int_0^T \langle C_{\theta_i}(\theta, S(\theta)) - f_{\theta_i}(\theta), p\rangle_{V^*,V} dt\,. \end{aligned} \tag{15}$$

To calculate the Hessian-vector product $\nabla^2\mathbf{j}(\theta)\zeta$ for some $\zeta \in \mathbb{R}^{n_\theta}$, we apply the same procedure to the auxiliary minimization problem $\min_{\theta\in\Theta} \nabla\mathbf{j}(\theta)^T\zeta$, which is then equivalent to

$$\begin{aligned} &\min_{\theta\in\Theta,(u,p)\in W(0,T)^2} \tilde{\mathbf{J}}(\theta,(u,p)) \\ &\text{s.t. } u_t + C(\theta,u) = f(\theta) \qquad u(0) = u_0 \\ &p_t - C_u(\theta,u)^* p = \mathbf{J}_u(\theta,u) \qquad p(T) = 0 \end{aligned} \tag{16}$$

with

$$\tilde{\mathbf{J}}(\theta,(u,p)) = \zeta^T\nabla_\theta\mathbf{J}(\theta,u) + \int_0^T \langle\zeta^T\nabla_\theta C(\theta,u) - \zeta^T\nabla_\theta f(\theta), p\rangle_{V^*,V} dt\,.$$

Defining v, w as the solutions of

$$\begin{aligned} v_t + C_u(\theta,S(\theta))v &= -\zeta^T\nabla_\theta C(\theta,S(\theta)) + \zeta^T\nabla_\theta f(\theta) && v(0) = 0 \\ w_t - C_u(\theta,S(\theta))^* w &= (B(\theta,S(\theta),P(\theta))^* v + (\zeta^T\nabla_\theta C_u(\theta,S(\theta)))^* P(\theta) \\ &\quad + \zeta^T\nabla_\theta\mathbf{J}_u(\theta,S(\theta)) + \mathbf{J}_{uu}(\theta,S(\theta))^* v && w(T) = 0 \end{aligned} \tag{17}$$

where we define B by

$$\langle B(\theta,a,b)c,d\rangle_{V^*,V} = \langle C_{uu}(\theta,a)(c,d),b\rangle_{V^*,V} \text{ for all } \theta\in\Theta\,,\ a,b,c,d\in V$$

and $P(\theta) = p$ as the solution to (14), we arrive at

$$\begin{aligned} (\nabla^2_\theta\mathbf{j}(\theta)\zeta)_i &= \frac{\partial}{\partial\theta_i}\tilde{\mathbf{j}}(\theta) = \frac{\partial}{\partial\theta_i}\tilde{\mathbf{J}}(\theta,(S(\theta),P(\theta))) \\ &= \frac{\partial}{\partial\theta_i}\Big(\zeta^T\nabla_\theta\mathbf{J}(\theta,S(\theta)) + \int_0^T \big\langle\zeta^T\nabla_\theta C(\theta,S(\theta)) - \zeta^T\nabla_\theta f(\theta), P(\theta)\big\rangle_{V^*,V} dt \\ &\qquad + \int_0^T \langle S(\theta)_t + C(\theta,S(\theta)) - f(\theta), w\rangle_{V^*,V} dt \\ &\qquad + \int_0^T \langle -P(\theta)_t + C_u(\theta,S(\theta))^* P(\theta) + \mathbf{J}_u(\theta,S(\theta)), v\rangle_{V^*,V} dt\Big) \\ &= (\nabla^2_\theta\mathbf{J}(\theta,S(\theta))\zeta)_i + \mathbf{J}_{\theta_i u}(\theta,S(\theta))v + \int_0^T \Big(\langle C_{\theta_i}(\theta,S(\theta)) - f_{\theta_i}(\theta), w\rangle_{V^*,V} \\ &\qquad + \langle((\nabla^2_\theta C(\theta,S(\theta)) - \nabla^2_\theta f(\theta))\zeta)_i + C_{\theta_i u}(\theta,S(\theta))v, P(\theta)\rangle_{V^*,V}\Big) dt\,. \end{aligned} \tag{18}$$

The numerical simulation of (12) with explicit or implicit time stepping can introduce numerical errors, which results in a divergence of the trajectory from the profile path and leads to an underestimation of the profile likelihood. This effect can be counterbalanced by the incorporation of a retraction term, which results in a minimization of $\mathbf{j}(\theta)$ for the given constraint,

$$\begin{pmatrix} \nabla^2_\theta\mathbf{j}(\theta_c) + \lambda_c\nabla^2_\theta\mathbf{g}(\theta_c) & \nabla_\theta\mathbf{g}(\theta_c) \\ \nabla_\theta\mathbf{g}(\theta_c)^T & 0 \end{pmatrix} \begin{pmatrix} \dot\theta_c \\ \dot\lambda_c \end{pmatrix} = \begin{pmatrix} -\gamma\nabla_\theta\mathbf{j}(\theta_c) \\ 1 \end{pmatrix} \tag{19}$$

with retraction factor $\gamma > 0$. This approach has been introduced by Chen and Jennrich [5]. Furthermore, to circumvent the potentially time-consuming calculation of the term $\nabla^2_\theta\mathbf{j}(\theta_c) + \lambda_c\nabla^2_\theta\mathbf{g}(\theta_c)$ they replace it with a positive definite matrix $\mathbf{w}(\theta_c)$, which depends at most on the first order derivatives of the parameter-to-state map. A possible choice for $\mathbf{w}(\theta_c)$ is the Fisher Information Matrix (FIM), which is for the objective function (5) given as

$$\begin{aligned} \mathbf{w}_{i,j}(\theta_c) &= \mathbf{J}_{\theta_i\theta_j}(\theta_c,S(\theta_c)) + \mathbf{J}_{\theta_i u}(\theta_c,S(\theta_c))S_{\theta_j}(\theta_c) + \mathbf{J}_{u\theta_j}(\theta_c,S(\theta_c))S_{\theta_i}(\theta_c) \\ &\quad + \mathbf{J}_{uu}(\theta_c,S(\theta_c))S_{\theta_i}(\theta_c)S_{\theta_j}(\theta_c). \end{aligned}$$

The replacement introduces an approximation error which also results in an underestimation of the profile likelihood. Chen and Jennrich [5] proved that as the retraction factor $\gamma > 0$ increases, the trajectory of (19) approaches the trajectory of (12). For $\gamma \to \infty$, we obtain the singular perturbed system which evolves along the profile [22]. In this reduced setting, the result from [5] applies directly.

5 Profile likelihood calculation for the full problem

In the previous section, the reduced problem was considered using the parameter-to-state map $S(\theta)$. The evaluation of $S(\theta)$ requires the accurate numerical simulation of the dynamical system. As this might be computationally inefficient, we introduce optimization and integration based profile likelihood calculation for the non-reduced form of the PDE constrained optimization problem.

5.1 Optimization based profile likelihood calculation

The optimization based profile likelihood calculation introduced in Section 4.1 relies on the solution of the reduced optimization problem (10) for every grid point c_l. The reduced optimization problem, however, can be replaced by the solution of the PDE constrained optimization problem,

$$\begin{aligned} \min_{\theta\in\Theta, u\in W(0,T)} \quad & \mathbf{J}(\theta,u) \\ \text{s.t. } \quad & u_t + C(\theta,u) = f(\theta) \\ & u(0) = u_0 \\ & G(\theta,u) = c. \end{aligned} \tag{20}$$

We denote the optimal solution by $(\theta_c, u_c) := \left(\hat{\theta}(c), \hat{u}(c)\right)$. This problem can be solved using local optimization, starting at initial points constructed from the previous grid points. For θ_c and u_c, similar extrapolation schemes can be used as for the reduced form.

5.2 Integration based profile likelihood calculation

For the derivation of the integration based profile likelihood calculation, we consider the Lagrange function of the PDE constrained optimization problem (20),

$$\tilde{\mathcal{L}}(\theta,u,p,\lambda) = \mathcal{L}(\theta,u,p) + \lambda(\mathbf{G}(\theta,u) - c) \tag{21}$$

with

$$\mathcal{L}(\theta,u,p) = \mathbf{J}(\theta,u) + \int_0^T (-\langle u, p_t\rangle_{V,V^*} + \langle C(\theta,u) - f(\theta), p\rangle_{V^*,V})dt.$$

and Lagrange multipliers λ and p. The first order optimality conditions for (20) at a minimizer (θ_c, u_c) are

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta_c,u_c,p_c) + \lambda_c \nabla_\theta \mathbf{G}(\theta_c,u_c) &= 0 \\ \nabla_u \mathcal{L}(\theta_c,u_c,p_c) + \lambda_c \nabla_u \mathbf{G}(\theta_c,u_c) &= 0 \\ \nabla_p \mathcal{L}(\theta_c,u_c,p_c) &= 0 \\ \mathbf{G}(\theta_c,u_c) &= c. \end{aligned} \tag{22}$$

The second line is the adjoint equation

$$\begin{cases} p_t - C_u(\theta,u)^* p = \mathbf{J}_u(\theta,u) \\ p(T) = 0 \end{cases} \tag{23}$$

and the third line is the state equation (1) for $p = p_c$, $u = u_c$ and $\theta = \theta_c$. Differentiating (22) with respect to c yields the following system for the evolution of $(\theta_c, u_c, p_c, \lambda_c)$:

$$\underbrace{\begin{pmatrix} \nabla_\theta^2 \mathcal{L} + \lambda_c \nabla_\theta^2 \mathbf{G} & \nabla_u \nabla_\theta \mathcal{L} + \lambda_c \nabla_u \nabla_\theta \mathbf{G} & \nabla_p \nabla_\theta \mathcal{L} & \nabla_\theta \mathbf{G} \\ \nabla_\theta \nabla_u \mathcal{L} + \lambda_c \nabla_\theta \nabla_u \mathbf{G} & \nabla_u^2 \mathcal{L} + \lambda_c \nabla_u^2 \mathbf{G} & \nabla_p \nabla_u \mathcal{L} & \nabla_u \mathbf{G} \\ \nabla_\theta \nabla_p \mathcal{L} & \nabla_u \nabla_p \mathcal{L} & 0 & 0 \\ \nabla_\theta \mathbf{G}^T & \nabla_u \mathbf{G}^T & 0 & 0 \end{pmatrix}}_{M_{\text{full}}(\theta_c,u_c,p_c,\lambda_c)} \begin{pmatrix} \dot{\theta}_c \\ \dot{u}_c \\ \dot{p}_c \\ \dot{\lambda}_c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{24}$$

where we skipped the arguments (θ_c, u_c, p_c) of the derivatives of $\mathcal{L}$ and $\mathbf{G}$. The derivatives of $\mathcal{L}$ are

$$\begin{aligned}
\mathcal{L}_{\theta_i\theta_j}(\theta_c, u_c, p_c) &= \mathbf{J}_{\theta_i\theta_j}(\theta_c, u_c) + \int_0^T \left\langle C_{\theta_i\theta_j}(\theta_c, u_c) - f_{\theta_i\theta_j}(\theta_c), p_c \right\rangle_{V^*,V} dt \\
\nabla_u \mathcal{L}_{\theta_i}(\theta_c, u_c, p_c) &= \mathbf{J}_{\theta_i u}(\theta_c, u_c) + C_{\theta_i u}(\theta_c, u_c)^* p_c \\
\nabla_p \mathcal{L}_{\theta_i}(\theta_c, u_c, p_c) &= C_{\theta_i}(\theta_c, u_c) - f_{\theta_i}(\theta_c) \\
\nabla_u^2 \mathcal{L}(\theta_c, u_c, p_c) &= \mathbf{J}_{uu}(\theta_c, u_c) + C_{uu}(\theta_c, u_c)^* p_c \\
\nabla_u \nabla_p \mathcal{L}(\theta_c, u_c, p_c) &= \partial_t + C_u(\theta_c, u_c)
\end{aligned} \tag{25}$$

The other mixed partial derivatives of the Langrange function follow by symmetry if all involved functions are twice continuously differentiable.

The trajectories of (24) provide the profile likelihood for the non-reduced problem, namely θ_c and u_c. For the numerical integration an explicit or implicit time stepping scheme can be used. Similarly to the reduced problem, the approximation errors can be reduced by introduction of a retraction term,

$$\begin{pmatrix} W_{uu\,c} & W_{u\theta\,c} & W_{pu\,c} & \nabla_u \mathbf{G} \\ W_{\theta u\,c} & W_{uu\,c} & W_{pu\,c} & \nabla_u \mathbf{G} \\ W_{\theta p\,c} & W_{up\,c} & W_{pp\,c} & 0 \\ \nabla_\theta \mathbf{G}^T & \nabla_u \mathbf{G}^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{\hat{\theta}}_c \\ \dot{\hat{u}}_c \\ \dot{\hat{p}}_c \\ \dot{\hat{\lambda}}_c \end{pmatrix} = \begin{pmatrix} -\gamma \nabla_\theta \mathcal{L} \\ -\gamma \nabla_u \mathcal{L} \\ -\gamma \nabla_p \mathcal{L} \\ 1 \end{pmatrix}, \tag{26}$$

with retraction factor $\gamma > 0$. The retraction damps the accumulation of numerical errors and ensures a more accurate profile likelihood approximation. Furthermore, the retraction allows for the replacement of the matrix $M_{\text{full}}(\theta_c, u_c, p_c, \lambda_c)$ with a positive definite matrix to circumvent the need for second order information. The resulting approximation error can be controlled using γ. A large retraction factor results, however, in an increased stiffness of the dynamical system. Extending the proof in [5] from finite dimensions and ODEs to the PDE setting in function spaces we can show that the difference between solutions to the original system and the approximated one with retraction can be made arbitrarily small by an appropriate choice of the retraction factor λ_c, see Proposition 5.1 below. For this purpose we abbreviate $\xi_c = (\theta_c, u_c, p_c)$, $\hat{\xi}_c = (\hat{\theta}_c, \hat{u}_c, \hat{p}_c)$, $\mathbf{G}(\xi) = \mathbf{G}(\theta, u)$, $X = \mathbb{R}^{n_\theta} \times W(0,T)^2$, so that we can rewrite (24) and (26) more compactly as

$$\begin{pmatrix} \nabla_\xi^2 \mathcal{L}(\xi_c) + \lambda_c \nabla_\xi^2 \mathbf{G}(\xi_c) & \nabla_\xi \mathbf{G}(\xi_c) \\ \nabla_\xi \mathbf{G}(\xi_c)^T & 0 \end{pmatrix} \begin{pmatrix} \dot{\xi}_c \\ \dot{\lambda}_c \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{27}$$

and

$$\begin{pmatrix} W_c & \nabla_\xi \mathbf{G}(\hat{\xi}_c) \\ \nabla_\xi \mathbf{G}(\hat{\xi}_c)^T & 0 \end{pmatrix} \begin{pmatrix} \dot{\hat{\xi}}_c \\ \dot{\hat{\lambda}}_c \end{pmatrix} = \begin{pmatrix} -\gamma \nabla_\xi \mathcal{L}(\hat{\xi}_c) \\ 1 \end{pmatrix}. \tag{28}$$

We assume that the family of linear operators $W_c : X \to X^*$ satisfies the following properties:

$$\forall c \in [c_0, c_1]\ \forall \xi, \zeta \in X : \langle W_c \xi, \zeta \rangle_{X^*,X} = \langle W_c \zeta, \xi \rangle_{X^*,X} \quad \text{(symmetry)} \tag{29}$$

$$\forall c \in [c_0, c_1]\ \forall \xi \in X : \langle W_c \xi, \xi \rangle_{X^*,X} \geq \gamma_W \|\xi\|_X^2 \quad \text{(positivity)} \tag{30}$$

$$\exists \bar{M} > 0\ \forall c \in [c_0, c_1] :\ \|W_c\| = \sup_{\xi,\zeta \in X,\, \|\xi\|_X \leq 1,\, \|\zeta\|_X \leq 1} \langle W_c \xi, \zeta \rangle_{X^*,X} \leq \bar{M}_W \quad \text{(boundedness)} \tag{31}$$

as well as the following sufficient second order condition at the minimizers (ξ_c, λ_c)

$$\exists \gamma_L > 0\, \forall c \in [c_0, c_1]\, \forall \zeta \in \nabla_\xi \mathbf{G}(\xi_c)_\perp :\ \langle (\nabla_\xi^2 \mathcal{L}(\xi_c) + \lambda_c \nabla_\xi^2 \mathbf{G}(\xi_c))\zeta, \zeta \rangle_{X^*,X} \geq \gamma_L \|\zeta\|_X^2 \tag{32}$$

where $\nabla_\xi \mathbf{G}(\xi_c)_\perp = \{\zeta \in X : \langle \nabla_\xi \mathbf{G}(\xi_c), \zeta \rangle_{X^*,X} = 0\}$ is the tangential cone corresponding to the equality constraint $\mathbf{G}(\xi) = 0$.

Proposition 5.1. *Let* $\mathbf{G}$, $\mathcal{L}$ *be twice continuously differentiable, let (29)–(32) be satisfied and let* ξ_c, $\hat{\xi}_c$, $c \in [c_0, c_1]$ *be solutions to (27) and (28), respectively.*
Then for any $\kappa > 0$, *and for any* $\tilde{\epsilon} > 0$ *sufficiently small, there exists* $\rho > 0$ *sufficiently small and* $\lambda > 0$ *sufficiently large, such that if* $e_{c_0} < \rho$ *then*

$$\forall c \in [c_0, c_1] \ \|\hat{\xi}_c - \xi_c\|_X \leq \rho \text{ and } e_c \leq \frac{\tilde{\epsilon}}{\kappa} + e_{c_0} \exp(-\kappa(c - c_0)) \tag{33}$$

holds, where $e_c = \langle W_c \hat{\xi}_c - \xi_c, \hat{\xi}_c - \xi_c \rangle_{X^*,X} \geq \gamma_W \|\hat{\xi}_c - \xi_c\|_X^2$.
Moreover, for any $\varepsilon > 0$ *and any* $\tilde{c} \in (c_0, c_1]$ *there exists* $\lambda > 0$ *such that*

$$\forall c \in [\tilde{c}, c_1] \ \|\hat{\xi}_c - \xi_c\|_X \leq \varepsilon \,. \tag{34}$$

The proof (see the Appendix) shows that $\lambda = \lambda_c$ can be chosen adaptively, depending on the artificial time parameter c.

6 Comparison of full and reduced formulation of integration based profile likelihood calculation

The full and reduced formulations of integration based profile calculation provide different view points on the problem. In the following, we will establish equivalence under the assumption of the identity $u_c = S(\theta_c)$. In addition, the computational implementation will be discussed.

6.1 Equivalence of calculated profile likelihoods

The reduced formulation (12) and the full formulation (24),

$$M_{\text{red}}(\theta_c, \lambda_c) \begin{pmatrix} \dot{\theta}_c \\ \dot{\lambda}_c \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad M_{\text{full}}(\theta_c, u_c, p_c, \lambda_c) \begin{pmatrix} \dot{\theta}_c \\ \dot{u}_c \\ \dot{p}_c \\ \dot{\lambda}_c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

provide two alternative approaches to calculate the profile likelihood path θ_c. The validity of the state equation for u_c, which is ensured by the initial condition satisfying (22) gives the identity $u_c = S(\theta_c)$. With this identity as well as the evolution (24), we can show the equivalence of both approaches.

Proposition 6.1. *Under Assumptions 2.1 and 4.1 solving the full system (24) and the reduced system (12) yields the same profile likelihood path* θ_c.

The proof of equivalence of (12) and (24) is provided in the Appendix. From the equivalence, we conclude that also the stabilized versions (19) and (26) yield the same results in the absence of numerical integration errors.

6.2 Implementation and computational properties

In the previous section we established equivalence of θ_c for (12) and (24). This equivalence is however not ensured for the result of the trajectories of (12) and (24) (or their stabilized versions (19) and (26)) obtained by numerical simulation. The implementation and computational requirements of full and reduced systems are considerably different.

For the numerical simulation of the reduced system, the matrix-vector product $M_{\text{red}}(\theta_c)(\dot{\theta}_c, \dot{\lambda}_c)^T$ has to be evaluated. This requires the numerical simulation of the model (2.1) for every point $(\theta_c, \lambda_c)^T$ and either $n_\theta(n_\theta + 3)/2$ linear forward PDE solves or three (one forward, two backward) linear PDE

solves (see Section 4.2). This implicit numerical simulation can exploit sophisticated numerical solvers, requires minimal storage but can be computationally demanding. In contrast, the full system provides an explicit form. When applying an iterative solver, e.g. a CG method, only matrix vector products are needed. Applying $M_{\text{full}}(\theta_c, u_c, p_c, \lambda_c)$ to a vector $(\dot{\theta}_c, \dot{u}_c, \dot{p}_c, \dot{\lambda}_c)^T$ only requires the evaluation of the linearization of the differential operator $\partial_t + C(\theta_c, .)$ and its adjoint, so no PDE solution. In the discretized setting, this amounts to evaluating some difference quotient with respect to time and to multiplying with the stiffness matrix in case of a parabolic PDE. The discretization of u_c and p_c in space and time can however require significant storage.

The mass matrices $M_{\text{red}}(\theta_c, \lambda_c)$ and $M_{\text{full}}(\theta_c, \lambda_c)$ can be singular. This happens, among others, in the presence of structural non-identifiabilities. For singular mass matrices, (12) and (24) (or their stabilized versions (19) and (26)) are differential algebraic equations and partial differential algebraic equations, respectively. The differentiation indexes are potentially unknown and might even be non-constant. Accordingly, sophisticated numerical methods are required (see, [3, 6] and references therein). Alternatively, pseudo-inverses of $M_{\text{red}}(\theta_c, \lambda_c)$ and $M_{\text{full}}(\theta_c, \lambda_c)$ [2] might be employed to derive approximating ODE and PDE systems.

In this paper the stabilized reduced system (19) is implemented. For the numerical simulation different adaptive solvers are employed.

7 Numerical evaluation of integration based profile likelihood calculation

In the following, we will illustrate the properties of the proposed integration based profile likelihood calculation method. For this purpose, we study a biological application, i.e. a PDE model for gradient formation. We consider a realistic measurement set-up but use artificial experimental data. This enables the comparison of the methods with the ground-truth available.

7.1 Mathematical model for gradient formation in fission yeast

To assess the properties of the proposed approach, we consider a model for gradient formation in fission yeast. Fission yeast cells are rod-shaped and their division is controlled by a gradient of the protein Pom1p in the cell membrane. Hersch et al. [14] modelled the dynamics of the concentration of Pom1p at a position x, $u(t, x)$ with units $\#/\mu m$, by

$$\begin{aligned} u_t &= Du_{xx} - \alpha u^2 + \frac{\beta}{\sqrt{2\pi}\rho} e^{-x^2/2\rho^2} && \text{for }]0, T[\times] - L, L[\\ \frac{\partial u}{\partial \nu} &= 0 && \text{for }]0, T[\times\{-L, L\} \\ u &= 0 && \text{for } \{t = 0\}\times] - L, L[\end{aligned} \tag{35}$$

with diffusion coefficient D, dimerisation rate α, influx rate β and source width ρ. The length along the membrane from the tip of the cell to the center is denoted by L. Model (35) meets Assumption 2.1.

We implemented the method of lines for model (35) in MATLAB. The system of ODEs was implemented in AMICI (*https://github.com/ICB-DCM/AMICI*) [21]. AMICI generates the first and second order sensitivity equations, enabling the evaluation of the Hessian and the Fisher Information Matrix. For the simulation, AMICI exploits the SUNDIALS solver suite [15].

7.2 Artificial experimental data

For a realistic evaluation of different profile likelihood calculation methods, we generated artificial experimental data similar to previously published datasets for the considered process (see [30]). Our artificial dataset consists of three individual datasets:

Table 1: **Parameter values and confidence intervals.** The parameter estimates are obtained using multi-start local optimization. The confidence intervals to a confidence level of 95% are computed with the integration based profile likelihood calculation in ODE formulation with the Hessian. For confidence intervals which extend beyond the considered parameter region we write $>$ or $<$ bound, indicating practical non-identifiability. All quantities are provided in seconds s, micrometer μm, number of molecules # and unit of fluorescence intensity ui.

parameter names	true value θ	estimated value $\widehat{\theta}$	95% confidence interval		units
			lower bound	upper bound	
D	0.10	0.15	0	> 12	$\mu m^2/s$
α	4.00×10^{-4}	6.07×10^{-4}	4.60×10^{-5}	> 1	$\mu m^3/(\# \cdot s)$
β	8.00×10^{3}	1.21×10^{4}	9.99×10^{2}	$> 5 \times 10^{8}$	$\#\mu m/s$
ρ	0.60	0.60	0.38	0.77	μm
s_1	2.87×10^{-4}	2.95×10^{-4}	2.45×10^{-4}	3.47×10^{-4}	$ui/\#$
s_2	2.70×10^{-5}	2.72×10^{-5}	2.26×10^{-5}	3.34×10^{-5}	$ui/\#$

- *Concentration profile:* The concentration profile provides the relative abundance of the signalling molecule along the membrane at time $t = 100\,s$. The interval $]-L, L[$ is divided in 60 equally sized regions Ω_k, yielding the observation operators

$$Q_k(\theta, u) = s_1 \int_{\Omega_k} u(t = 100, x)dx \quad \text{for} \quad k = 1, \dots, 60,$$

  with scaling factor s_1 and region $\Omega_k = \left[-7 + \frac{7}{30}(k-1), -7 + \frac{7}{30}k\right]$, $k = 1, \dots, 60$.

- *Time course:* The time course data provide the scaled overall protein abundance at 10 equally spaced time points $t_k \in [0, 60]\,s$. The observation operators are

$$Q_{60+k}(\theta, u) = s_2 \int_{-L}^{L} u(t_k, x)dx \quad \text{for} \quad k = 1, \dots, 10,$$

  with scaling factor s_2.

- *Quantification:* The quantification provides the absolute abundance of the signalling molecule at time point $t = 100\,s$,

$$Q_{71}(\theta, u) = \int_{-L}^{L} u(t = 100, x)dx.$$

The artificial data sets were obtained by simulating model (35) for the parameters provided in Table 1, and subsequently adding normally distributed measurement noise. The final data set is the mean $\bar{y}_k$ and standard deviation σ_k, $k = 1, \dots, 71$, as depicted in Figure 3. In biological applications the acquisition of measurement data is often challenging. Instead of a single highly-informative experiment, merely a series of measurements with low information content can be performed. This commonly results in observation operators $Q(\theta, u)$ with a non-standard structure.

7.3 Parameter optimization

For the estimation of the unknown parameter vector $\theta = (D, \alpha, \beta, \rho, s_1, s_2)^T$ we use maximum likelihood estimation. The measurement noise is assumed to be normally distributed with the measured standard deviation σ_k (see error bars in Figure 3). This yields the reduced optimization problem

$$\min_{\theta \in \mathbb{R}^n} \mathbf{j}(\theta) = \frac{1}{2} \sum_{k=1}^{71} \left(\log\left(2\pi\sigma_k^2\right) + \left(\frac{\bar{y}_k - Q_k(\theta, S(\theta))}{\sigma_k} \right)^2 \right) \tag{36}$$

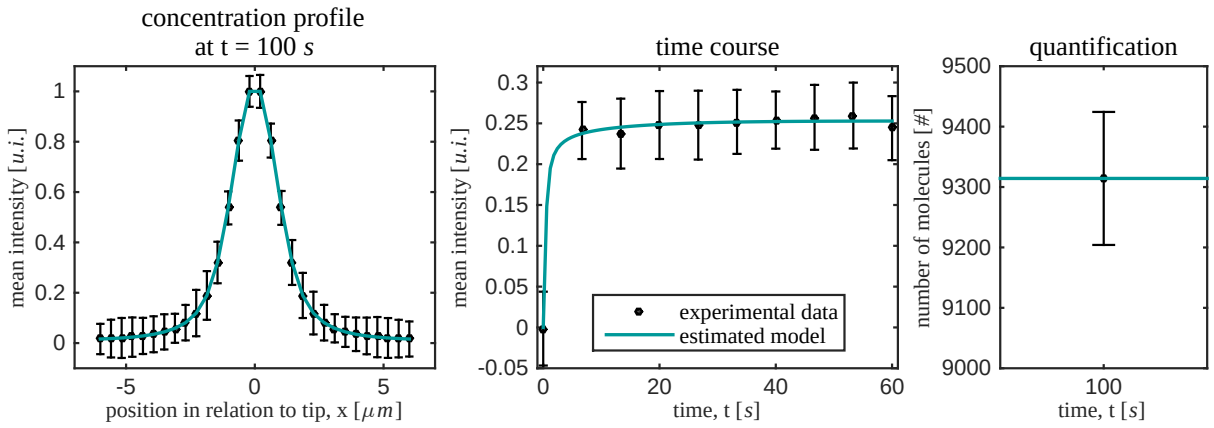

Figure 3: **Artificial experimental data and model fit.** The measurement data (•) and its standard deviation (error bar) are depicted along with the best model fit (▬) for (left) the concentration profile, (middle) the time course and (right) the protein abundance.

in which $S(\theta)$ denotes the parameter-to-state map of (35), which is evaluated using numerical integration. The optimum of (36) is determined using multi-start local optimization. Therefore, the MATLAB optimizer *fmincon* is initialised at 100 different starting points chosen with a space filling design, i.e. latin hyper cube sampling. More than 90% of these optimizations converged to the same optimal likelihood value, which we assume to be global. The estimation results are shown in Table 1 and Figure 3.

7.4 Profile likelihood calculation

To assess the uncertainty of the parameter estimate, we compute the profile likelihoods $\mathrm{PL}_{\theta_i}(c)$, $i = 1, \ldots, 6$. Therefore we employ existing

- optimization based profile likelihood calculation with 0th and 1st order proposal,

as well as the proposed

- integration based profile likelihood calculation with Hessian and
- integration based profile likelihood calculation with FIM.

For the integration based methods we compare the numerical implementation as DAE system (19) and as ODE system,

$$\begin{pmatrix} \dot{\theta}_c \\ \dot{\lambda}_c \end{pmatrix} = M_{\mathrm{red}}(\theta_c, \lambda_c)^+ \begin{pmatrix} \nabla_\theta \mathbf{j}(\theta_c) \\ 1 \end{pmatrix},$$

with $M_{\mathrm{red}}(\theta_c, \lambda_c)^+$ denoting the Moore-Penrose pseudo-inverse of $M_{\mathrm{red}}(\theta_c, \lambda_c)$. The use of the Moore-Penrose pseudo-inverse is necessary as $M_{\mathrm{red}}(\theta_c, \lambda_c)$ is singular in some of the considered situations.

All methods were implemented in the open-source Parameter EStimation TOolbox (PESTO) for MATLAB (*https://github.com/ICB-DCM/PESTO*). The optimization based calculation exploits the MATLAB function *fmincon* with a user supplied gradient of the objective function. The integration based calculation is implemented using the MATLAB function *ode15s*, which is suited for ODEs and DAEs.

The profile likelihood calculation using the aforementioned methods provided consistent results. Optimization based profile likelihood calculation, integration based profile likelihood calculation using the Hessian and integration based profile likelihood calculation using the FIM (with retraction factor $\gamma > 2$) indicate that for the given data set ρ, s_1 and s_2 are practically identifiable for a confidence

level of 95% while D, α and β are practically non-identifiable (see Table 3 and Figure 4A). For $\gamma < 6$, integration based profile likelihood calculation using the FIM yielded an underestimation of the profile likelihood. This effect is worse for practically non-identifiable parameters than for practically identifiable parameters. However, for increasing values of γ the profile likelihood converged to the profile likelihood obtained with the optimization based method (Figure 4A (right)). Integration based profile likelihood calculation using the Hessian provided accurate results independent of γ. Differences in the path θ_c resulting from the implementation of DAE or ODE were negligible.

The comparison of the computation time for the different methods revealed that the optimization based profile calculation with 1st order proposal is substantially faster than optimization based profile calculation with 0th order proposal. The use of the 1st order proposal reduces the number of grid points, i.e the cardinality of $\{c_l\}_l$. In addition, the starting points $\theta_{c_l}^{(0)}$ are on average slightly closer to the constraint optimal values θ_{c_l}, reducing the computation time required for local optimization. Despite the efficiency of optimization based methods with 1st order proposal, these methods are outperformed by integration based profile likelihood calculation using the Hessian. For the practically identifiable parameter ρ the implementation of integration based profile likelihood calculation using the Hessian as ODE results in a speed-up by a factor of five compared to optimization based profile calculation with 0th order proposal (Figure 4A (left)). For the practically non-identifiable parameter α the efficiency improvement was a factor 144 compared to the optimization based methods with 0th order proposal and a factor five compared to optimization based methods with 1st order proposal. One reason for the improved computational efficiency was the adaptive choice of the evaluation points performed by the ODE solver, which allowed for larger steps in regions with smaller curvature. This effect was reduced for the FIM, as the retraction increases the stiffness of the DAE or ODE. The decrease in the step sizes due to the stiffness yielded more function evaluations and an increased computation time. Surprisingly, this increase also outweighed the higher computation cost of computing second order sensitivities. Furthermore our analysis of the computation times demonstrated that for this problem the ODE implementation was computationally more efficient than the DAE implementation.

Integration based profile calculation methods allow for the analysis of individual parameters $\mathbf{g}(\theta) = \theta_i$ but also for more complex expressions. We considered the parameter ratio $\mathbf{g}(\theta, u) = \frac{\alpha}{\beta}$. While the individual parameters possess an unbounded confidence interval and are practically non-identifiable, the ratio is practically identifiable and possesses a finite confidence interval (Figure 5A). This indicates that influx (related to β) and outflow (related to α) are balanced. In addition, the analysis of the quotient of the molecule abundance at the tip compared to the abundance at $x = 2$ for the time point $t = 100$, $\mathbf{G}(\theta, u) = u(t = 100, x = 2)/u(t = 100, x = 0)$, revealed that the steepness of the gradient is well determined (Figure 5B).

In summary, the numerical evaluation for Pom1p signalling revealed the accuracy and efficiency of the integration based profile likelihood calculation methods. Beyond individual parameters, integration based methods facilitated uncertainty analysis for a range of scalar model properties.

8 Conclusion

In this paper we considered profile likelihood methods for uncertainty analysis in PDE constrained inverse problems. Profile likelihoods provide statistically interpretable confidence bounds for parameters and model predictions using maximum projections of the likelihood. We formulated optimization based profile likelihood calculation methods for the reduced and the full problem. As optimization based approaches can become computationally demanding for PDE constrained problems, we extended the results for the reduced problem by Chen and Jennrich [5] to PDEs. In addition, we formulated an integration based profile likelihood calculation method for the full problem and established equivalence with the reduced formulation.

The integration based methods provide the exact profile likelihood, if second order information, i.e. Hessian-vector products, are available. We introduced an approximation of the integration based approach for the full problem to circumvent the second order information. A bound for the approximation

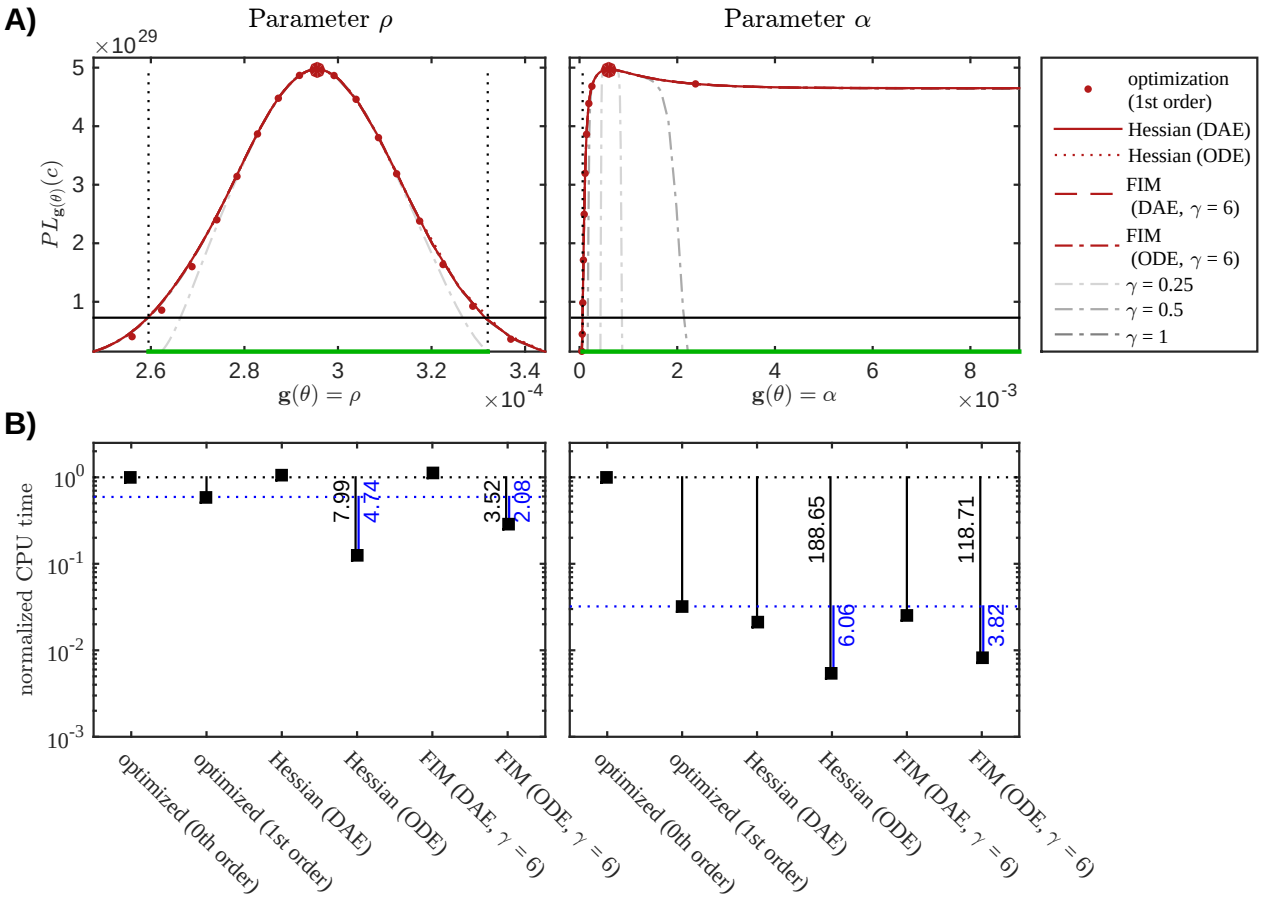

Figure 4: **Profile likelihood calculation for individual parameters θ_i.** (A) Profile likelihood for ρ and α calculated using the optimization based and integration based methods. For the practically identifiable parameter ρ (left), almost all profiles agree perfectly. For the practically non-identifiable parameter α (right), integration based methods using the Fisher Information Matrix and $\gamma < 2$ underestimate the profile likelihood. (B) Normalised CPU time for different profile likelihood calculation methods. The numbers indicate the speed up compared to the optimization based method with 0th order proposal (—) or the 1st order proposal (—).

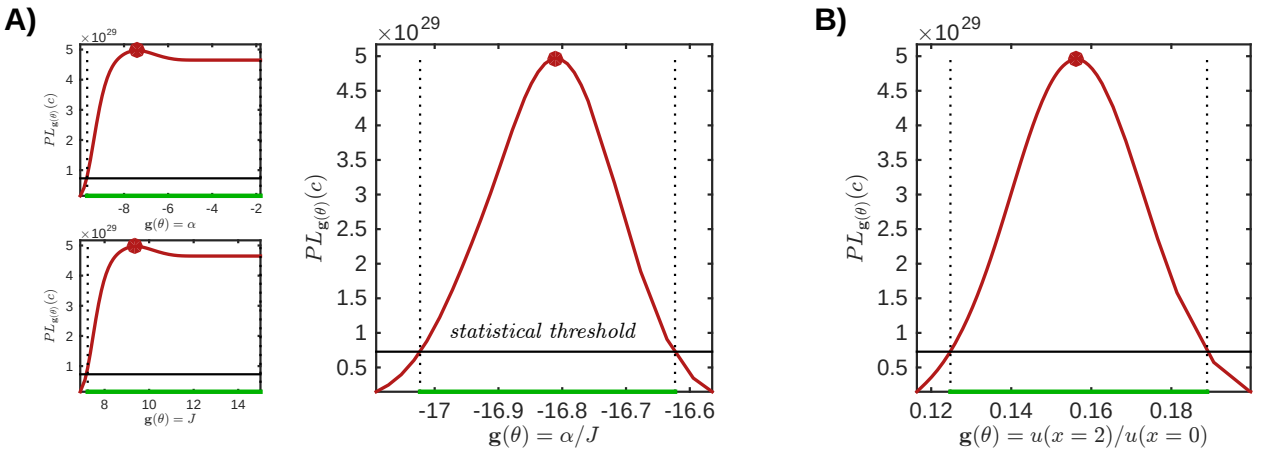

Figure 5: **Profile likelihood calculation for parameter ratio and model prediction.** (A) Profile likelihoods (—) for α and β revealing practical non-identifiability. Profile likelihood for the parameter combination $\mathbf{G}(\theta, u) = \frac{\alpha}{\beta}$ with a finite confidence interval $\mathrm{CI}_{0.95,\mathbf{g}(\theta)}$ (—). (B) Profile likelihood and confidence interval for $\mathbf{G}(\theta, u) = u(t = 100, x = 2)/u(t = 100, x = 0)$. All profile likelihoods are computed using the ODE formulation of the integration based method with the Hessian.

error was provided and convergence with respect to the retraction factor was established.

Optimization and integration based profile likelihood calculation methods for the reduced problem were assessed for a semi-linear PDE model of gradient formation in fission yeast. For both, practically identifiable and practically non-identifiable parameters, integration based methods were 4- to 5-fold faster than state-of-the-art optimization based approaches. The precise speed-up depended on the specific implementation, using exact second order information improved accuracy as well as computational efficiency. We note that for the practically non-identifiable parameters of the model, uncertainty analysis methods based on local approximations fail to provide realistic confidence bounds [26].

The implementation used for the comparison exploited forward methods for the calculation of gradient and Hessian of the objective function. We expect a further improvement of the computational efficiency by using adjoint methods. In addition, the implementation of integration based profile likelihood calculation for the full formulation promises an improved computational efficiency. The solution of the PDE model – currently performed in each solver step – would be circumvented and instead the nested ODE-PDE system (24) would be considered, which is an abstract ODE with respect to the artificial time c, where the operators are partial differential operators with respect to physical space and time. Matrix-vector products with $M_{\text{full}}(\theta_c, u_c, p_c, \lambda_c)$ from (24) will involve this application of partial differential operators, which makes clear that, e.g., CG type solution methods for such systems require proper preconditioning (see, e.g., [1, 31]). In addition, the implementation could be extended to the profile likelihood analysis of time-dependent properties (see [13]).

In summary, this study presented optimization and integration based profile likelihood calculation methods for PDE constrained problems. Besides exact methods, we present approximation and corresponding error bounds. The methods for the reduced formulation are implemented in the open-source MATLAB toolbox PESTO, which will facilitate the application of the method and simplify the development of extensions. This is particularly interesting as the methods we developed can be easily transferred to a broad class of PDE models. Accordingly, this study can help to improve uncertainty analysis in a number of scientific fields.

Appendix

Proof of Proposition 5.1

Assumptions (29) and (30) imply that $(\xi, \zeta)_{W_c} = \langle W_c \xi, \zeta \rangle_{X^*,X}$ defines an inner product on the space X.

For any $c \in [c_0, c_1]$, we abbreviate $\tilde{g}_c = W_c^{-1} \nabla_\xi \mathbf{G}(\hat{\xi}_c) \in X$ and $\tilde{l}_c = W_c^{-1} \nabla_\xi \mathcal{L}(\hat{\xi}_c) \in X$, and define the mapping $P_c : X \to X$ by

$$P_c \zeta = \frac{\langle \nabla_\xi \mathbf{G}(\hat{\xi}_c), \zeta \rangle_{X^*,X}}{\langle \nabla_\xi \mathbf{G}(\hat{\xi}_c), \tilde{g}_c \rangle_{X^*,X}} \tilde{g}_c$$

Note that P_c is linear and idempotent, i.e., a projection onto the one-dimensional linear space $\operatorname{span}(\tilde{g}_c)$, actually the orthogonal projection with respect to the inner product $(\cdot,\cdot)_{W_c}$. The second line in (28) yields

$$P \dot{\hat{\xi}}_c = \frac{1}{\langle \nabla_\xi \mathbf{G}(\hat{\xi}_c), \tilde{g}_c \rangle_{X^*,X}} \tilde{g}_c$$

and the first line in (28) (after application of W_c^{-1}) together with the fact that $(I - P_c)\tilde{g}_c = 0$ yields

$$(I - P) \dot{\hat{\xi}}_c = -\gamma (I - P_c) \tilde{l}_c$$

hence altogether we have eliminated $\hat{\lambda}_c$ and end up with the identity

$$\dot{\hat{\xi}}_c = \frac{1}{\langle \nabla_\xi \mathbf{G}(\hat{\xi}_c), \tilde{g}_c \rangle_{X^*,X}} \tilde{g}_c - \gamma (I - P_c) \tilde{l}_c \,.$$

i.e., the following evolution equation for the error:

$$\dot{\hat{\xi}}_c - \dot{\xi}_c = \frac{1}{\langle \nabla_\xi \mathbf{G}(\hat{\xi}_c), \tilde{g}_c\rangle_{X^*,X}} \tilde{g}_c - \dot{\xi}_c - \gamma(I - P_c)\tilde{l}_c\,. \tag{37}$$

Here, the last term is responsible for retraction. Indeed, using symmetry (29) and the definition of P_c, which yields, for any $\xi, \zeta \in X$

$$\langle W_c\xi, (I-P_c)\zeta\rangle_{X^*,X} = \langle W_c\zeta, (I-P_c)\xi\rangle_{X^*,X}$$

as well as

$$\langle W_c\tilde{g}_c, (I-P_c)\xi\rangle_{X^*,X} = \langle W_c\xi, (I-P_c)\tilde{g}_c\rangle_{X^*,X} = \langle W_c\xi, 0\rangle_{X^*,X} = 0 \tag{38}$$

we get

$$\begin{aligned}\langle W_c(\hat{\xi}_c - \xi_c), (I-P_c)\tilde{l}_c\rangle_{X^*,X} &= \langle W_c\tilde{l}_c, (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\\ &= \langle W_c(\tilde{l}_c + \lambda_c\tilde{g}_c), (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\\ &= \langle \nabla_\xi\mathcal{L}(\hat{\xi}_c) + \lambda_c\nabla_\xi\mathbf{G}(\hat{\xi}_c), (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\end{aligned}$$

Since $\nabla_\xi\mathcal{L}(\xi_c) + \lambda_c\nabla_\xi\mathbf{G}(\xi_c) = 0$, we have

$$\begin{aligned}&\langle \nabla_\xi\mathcal{L}(\hat{\xi}_c) + \lambda_c\nabla_\xi\mathbf{G}(\hat{\xi}_c), (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\\ &= \langle (\nabla^2_\xi\mathcal{L}(\xi_c) + \lambda_c\nabla^2_\xi\mathbf{G}(\xi_c))(\hat{\xi}_c - \xi_c), (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X} + \langle \mathrm{tay}, (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\end{aligned}$$

for

$$\mathrm{tay} = \nabla_\xi\Phi(\hat{\xi}_c) - \nabla_\xi\Phi(\xi_c) - \nabla^2_\xi\Phi(\xi_c)(\hat{\xi}_c - \xi_c) = o(\|\hat{\xi}_c - \xi_c\|_X)$$

where $\Phi(\xi) = \mathcal{L}(\xi) + \lambda_c\mathbf{G}(\xi)$. Moreover, $(I-P_c)(\hat{\xi}_c - \xi_c) \in \nabla_\xi\mathbf{G}(\xi_c)_\perp$ (cf. (38)), and by $\langle \nabla_\xi\mathbf{G}(\hat{\xi}_c), \hat{\xi}_c - \xi_c\rangle_{X^*,X} = \mathbf{G}(\hat{\xi}_c) - \mathbf{G}(\xi_c) + \frac{1}{2}\nabla^2_\xi\mathbf{G}(\hat{\xi}_c + \tau(\xi_c - \hat{\xi}_c))(\xi_c - \hat{\xi}_c)^2 = c - c + \frac{1}{2}\nabla^2_\xi\mathbf{G}(\hat{\xi}_c + \tau(\xi_c - \hat{\xi}_c))(\xi_c - \hat{\xi}_c)^2$ we have

$$\|P_c(\hat{\xi}_c - \xi_c)\|_X = \frac{|\langle \nabla_\xi\mathbf{G}(\hat{\xi}_c), \hat{\xi}_c - \xi_c\rangle_{X^*,X}|}{|\langle \nabla_\xi\mathbf{G}(\hat{\xi}_c), \tilde{g}_c\rangle_{X^*,X}|}\|\tilde{g}_c\|_X \le \bar{M}_c\,\|\hat{\xi}_c - \xi_c\|^2_X$$

for $\bar{M}_c = \frac{\sup_{\xi\in[\xi_c,\hat{\xi}_c]}\|\nabla^2_\xi\mathbf{G}(\xi)\|}{2\gamma_W\|\tilde{g}_c\|_X}$. Thus by (32) we can further estimate

$$\begin{aligned}&\langle W_c(\hat{\xi}_c - \xi_c), (I-P_c)\tilde{l}_c\rangle_{X^*,X}\\ &\ge \gamma_L\|(I-P_c)(\hat{\xi}_c - \xi_c)\|^2_X - \gamma_L\bar{M}_c^2\,\|\hat{\xi}_c - \xi_c\|^4_X + \langle \mathrm{tay}, (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\\ &\quad - \|\nabla^2_\xi\mathcal{L}(\xi_c) + \lambda_c\nabla^2_\xi\mathbf{G}(\xi_c)\|\,\bar{M}_c\,\|\hat{\xi}_c - \xi_c\|^2_X\|(I-P_c)(\hat{\xi}_c - \xi_c)\|_X\\ &\ge \frac{\gamma_L}{2}\|\hat{\xi}_c - \xi_c\|^2_X + \langle \mathrm{tay}, (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\\ &\quad - \|\nabla^2_\xi\mathcal{L}(\xi_c) + \lambda_c\nabla^2_\xi\mathbf{G}(\xi_c)\|\,\bar{M}_c\,\|\hat{\xi}_c - \xi_c\|^2_X\|(I-P_c)(\hat{\xi}_c - \xi_c)\|_X\end{aligned}$$

Thus applying $W_c(\hat{\xi}_c - \xi_c)$ to $\dot{\hat{\xi}}_c - \dot{\xi}_c$ and using the identity

$$\langle W_c(\hat{\xi}_c - \xi_c), \dot{\hat{\xi}}_c - \dot{\xi}_c\rangle_{X^*,X} = \frac{1}{2}\frac{d}{dc}\langle W_c(\hat{\xi}_c - \xi_c), \hat{\xi}_c - \xi_c\rangle_{X^*,X} - \frac{1}{2}\langle \dot{W}_c(\hat{\xi}_c - \xi_c), \hat{\xi}_c - \xi_c\rangle_{X^*,X}$$

that follows from symmetry (29), we get from (37)

$$\begin{aligned}&\frac{1}{2}\frac{d}{dc}\langle W_c(\hat{\xi}_c - \xi_c), \hat{\xi}_c - \xi_c\rangle_{X^*,X}\\ &\le -\gamma\frac{\gamma_L}{2}\|\hat{\xi}_c - \xi_c\|^2_X + \gamma\gamma_L\bar{M}_c^2\,\|\hat{\xi}_c - \xi_c\|^4_X - \gamma\langle \mathrm{tay}, (I-P_c)(\hat{\xi}_c - \xi_c)\rangle_{X^*,X}\\ &\quad + \gamma\|\nabla^2_\xi\mathcal{L}(\xi_c) + \lambda_c\nabla^2_\xi\mathbf{G}(\xi_c)\|\,\bar{M}_c\,\|\hat{\xi}_c - \xi_c\|^2_X\|(I-P_c)(\hat{\xi}_c - \xi_c)\|_X\\ &\quad + \frac{1}{2}\|\dot{W}_c\|\,\|\hat{\xi}_c - \xi_c\|^2_X + \left(\frac{1}{\|\tilde{g}_c\|_X} + \sqrt{\langle W_c\dot{\xi}_c, \dot{\xi}_c\rangle_{X^*,X}}\right)\sqrt{\langle W_c(\hat{\xi}_c - \xi_c), \hat{\xi}_c - \xi_c\rangle_{X^*,X}}\,,\end{aligned}$$

where by Young's inequality, the last term can be bounded by

$$\frac{\tilde{\epsilon}}{2} + \frac{1}{2\tilde{\epsilon}} \left(\frac{1}{\|\tilde{g}_c\|_X} + \sqrt{\langle W_c \dot{\xi}_c, \dot{\xi}_c \rangle_{X^*,X}} \right)^2 \langle W_c(\hat{\xi}_c - \xi_c), \hat{\xi}_c - \xi_c \rangle_{X^*,X} \,,$$

where $\tilde{\epsilon} > 0$ can still be chosen. Thus using (30), we end up with an estimate of the form

$$\dot{e}_c \leq \tilde{\epsilon} - \left(\gamma m - \gamma f(e_c) - M_c^{\tilde{\epsilon}} \right) e_c \tag{39}$$

for $e_c = \langle W_c(\hat{\xi}_c - \xi_c), \hat{\xi}_c - \xi_c \rangle_{X^*,X}$, where $m = \frac{\gamma_L}{M_W}$, $M_c^{\tilde{\epsilon}} = \frac{\|\dot{W}_c\|}{\gamma_W} + \frac{1}{\tilde{\epsilon}} \left(\frac{1}{\|\tilde{g}_c\|_X} + \sqrt{\langle W_c \dot{\xi}_c, \dot{\xi}_c \rangle_{X^*,X}} \right)^2$, and $f(t) = o(t)$ as $t \to 0$ and without loss of generality f is monotonically decreasing.
So there exists $\rho > 0$ such that $f(\rho) < \frac{m}{2}$. We impose the initial smallness $e_{c_0} < \rho$ and, for any $\tilde{\epsilon} \in]0, \frac{\rho - e_{c_0}}{c_1 - c_0}[$, choose $\gamma \geq \frac{2}{m} M_c^{\tilde{\epsilon}}$. With this choice we have, first of all, that $e_c \leq \rho$ for all $c \in [c_0, c_1]$, which can be seen as follows: Assume, on the contrary, that for some $c \in [c_0, c_1]$, $e_c > \rho$ holds and define c_2 to be the smallest such c, $c_2 = \inf\{c \in [c_0, c_1] \,:\, e_c > \rho\}$. Then by the initial smallness condition, c_2 must be strictly larger than c_0, by minimality of c_2 we have $e_c \leq \rho$ for all $c \in [c_0, c_2]$, and finally, by the sequential definition of the infimum we get $e_{c_2} \geq \rho$. Integration of (39) therefore by the choice of ρ and γ as well as the initial smallness condition yields

$$\begin{aligned} e_{c_2} &\leq e_{c_0} + \tilde{\epsilon}(c_2 - c_0) - \int_{c_0}^{c_2} \left(\gamma m - \gamma f(e_c) - M_c^{\tilde{\epsilon}} \right) e_c \, dc \\ &\leq e_{c_0} + \tilde{\epsilon}(c_2 - c_0) - \int_{c_0}^{c_2} \left(\gamma \frac{m}{2} - M_c^{\tilde{\epsilon}} \right) e_c \, dc \\ &\leq e_{c_0} + \tilde{\epsilon}(c_2 - c_0) < \rho \,, \end{aligned}$$

which contradicts $e_{c_2} \geq \rho$. Thus we have shown the boundedness estimate in (33), which additionally implies that $f(e_c) \leq \frac{m}{2}$ for all $c \in [c_0, c_2]$ and hence

$$\dot{e}_c \leq \tilde{\epsilon} - \left(\gamma \frac{m}{2} - M_c^{\tilde{\epsilon}} \right) e_c \tag{40}$$

To prove the exponential decay estimate in (33) for given $\kappa > 0$, $\tilde{\epsilon} \in]0, \frac{\rho - e_{c_0}}{c_1 - c_0}[$, we choose λ possibly larger, namely $\lambda \geq \frac{2}{m}(\kappa + M_c^{\tilde{\epsilon}})$ to obtain from (40)

$$\dot{e}_c \leq \tilde{\epsilon} - \kappa e_c \tag{41}$$

and Gronwall's inequality, applied to $e_c - \frac{\tilde{\epsilon}}{\kappa}$, that

$$e_c - \frac{\tilde{\epsilon}}{\kappa} \leq \left(e_{c_0} - \frac{\tilde{\epsilon}}{\kappa} \right) \exp(-\kappa(c - c_0)) \,.$$

Finally, (34) follows by choosing $\tilde{\epsilon} \leq \frac{\varepsilon}{2}$, $\kappa \geq \max\left\{1, \frac{\ln(2e_{c_0}) - \ln(\varepsilon)}{c - c_0}\right\}$, $\lambda \geq \frac{2}{m}(\kappa + M_c^{\tilde{\epsilon}})$.

Proof of Proposition 6.1

The operators $\nabla_u \nabla_p \mathcal{L}(\theta_c, u_c, p_c, \lambda_c)$ and $\nabla_p \nabla_u \mathcal{L}(\theta_c, u_c, p_c, \lambda_c)$ represent the linearised state and the adjoint equation, respectively and are thus invertible under Assumptions 2.1 and 4.1. Therefore we can formally eliminate the variables $(\dot{u}_c, \dot{p}_c)$ by means of the second and third line in the system (24), which yields

$$\begin{aligned} \dot{u}_c &= -(\nabla_u \nabla_p \mathcal{L})^{-1} \nabla_\theta \nabla_p \mathcal{L} \, \dot{\theta}_c \\ \dot{p}_c &= (\nabla_p \nabla_u \mathcal{L})^{-1} \Big(\nabla_u^2 \mathcal{L} (\nabla_u \nabla_p \mathcal{L})^{-1} \nabla_\theta \nabla_p \mathcal{L} - \nabla_\theta \nabla_u \mathcal{L} \Big) \dot{\theta}_c \,, \end{aligned} \tag{42}$$

where we have skipped the arguments (θ_c, u_c, p_c) of the Lagrangian for better readability. Inserting this into (24) yields

$$\begin{pmatrix} \tilde{M} + \lambda_c \nabla_\theta^2 g(\theta_c) & \nabla_\theta g(\theta_c) \\ \nabla_\theta g(\theta_c)^T & 0 \end{pmatrix} \begin{pmatrix} \dot{\theta}_c \\ \dot{\lambda}_c \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with

$$\begin{aligned} \tilde{M} = {} & \nabla_\theta^2 \mathcal{L} + \nabla_u \nabla_\theta \mathcal{L} (-\nabla_u \nabla_p \mathcal{L})^{-1} \nabla_\theta \nabla_p \mathcal{L} - \nabla_p \nabla_\theta \mathcal{L} (\nabla_p \nabla_u \mathcal{L})^{-1} \nabla_\theta \nabla_u \mathcal{L} \\ & - \nabla_p \nabla_\theta \mathcal{L} (\nabla_p \nabla_u \mathcal{L})^{-1} \nabla_u^2 \mathcal{L} (-\nabla_u \nabla_p \mathcal{L})^{-1} \nabla_\theta \nabla_p \mathcal{L} \,. \end{aligned}$$

Thus to show equivalence with (12) it only remains to verify that $\tilde{M} = \nabla_\theta^2 \mathbf{j}(\theta_c)$. With the second derivatives according to (25) and $(-\nabla_p \mathcal{L}_{\theta_i})(\nabla_p \nabla_u \mathcal{L})^{-1} = S_{\theta_i}(\theta_c)$ we get

$$\begin{aligned} \tilde{M}_{i,j} = {} & \mathbf{J}_{\theta_i \theta_j}(\theta_c, S(\theta_c)) + \int_0^T \left\langle C_{\theta_i \theta_j}(\theta_c, S(\theta_c)) - f_{\theta_i \theta_j}(\theta_c), p_c \right\rangle_{V^*,V} dt + \mathbf{J}_{\theta_i u}(\theta_c, S(\theta_c)) S_{\theta_j}(\theta_c) \\ & + \int_0^T \left\langle C_{\theta_i u}(\theta_c, S(\theta_c)) S_{\theta_j}(\theta_c), p_c \right\rangle_{V^*,V} dt + \int_0^T \left\langle C_{uu}(\theta_c, S(\theta_c)) S_{\theta_i}(\theta_c) S_{\theta_j}(\theta_c), p_c \right\rangle_{V^*,V} dt \\ & + \mathbf{J}_{uu}(\theta_c, S(\theta_c)) S_{\theta_i}(\theta_c) S_{\theta_j}(\theta_c) + \mathbf{J}_{u\theta_j}(\theta_c, S(\theta_c)) S_{\theta_i}(\theta_c) + \int_0^T \left\langle C_{u\theta_j}(\theta_c, S(\theta_c)) S_{\theta_i}(\theta_c), p_c \right\rangle_{V^*,V} dt \,. \end{aligned}$$

Using the definition of $h_{ij}^{\theta_c}$ and the identity

$$\int_0^T \left\langle h_{ij}^{\theta_c}, p_c \right\rangle_{V^*,V} dt = \mathbf{J}_u(\theta_c, S(\theta_c)) S_{\theta_i \theta_j}(\theta_c)$$

we obtain

$$\begin{aligned} \tilde{M}_{i,j} = {} & \mathbf{J}_{\theta_i \theta_j}(\theta_c, S(\theta_c)) + \mathbf{J}_{\theta_i u}(\theta_c, S(\theta_c)) S_{\theta_j}(\theta_c) + \mathbf{J}_{uu}(\theta_c, S(\theta_c)) S_{\theta_i}(\theta_c) S_{\theta_j}(\theta_c) \\ & + \mathbf{J}_{u\theta_j}(\theta_c, S(\theta_c)) S_{\theta_i}(\theta_c) + \mathbf{J}_u(\theta_c, S(\theta_c)) S_{\theta_i \theta_j}(\theta_c) = \mathbf{j}_{\theta_i \theta_j}(\theta_c) \end{aligned}$$

which establishes equivalence.

Acknowledgments

The fourth author acknowledges support by the Austrian Science Fund FWF under grant I2271. The first and fourth author were supported by the Karl Popper Kolleg "Modeling-Simulation-Optimization" funded by the Alpen-Adria-Universität Klagenfurt and by the Carinthian Economic Promotion Fund (KWF)
Moreover, we wish to thank both reviewers for fruitful comments leading to an improved version of the manuscript.

References

[1] A. Battermann and E. Sachs, *Block preconditioner for kkt systems in pde-governed optimal control problems*, in Workshop on Fast Solutions of Discretized Optimization Problems, R. Hoppe, K.-H. Hoffmann, and V. Schulz, eds., Birkhäuser, 2001, pp. 1–18.

[2] F. J. Beutler, *The operator theory of the pseudo-inverse*, J. Math. Anal. Appl., 10 (1965), pp. 451–470.

[3] P. N. Brown, A. C. Hindmarsh, and L. R. Petzold, *Using Krylov methods in the solution of large-scale differential-algebraic systems*, SIAM J. Sci. Comput., 15 (1994), pp. 1467–1488.

[4] J.-S. Chen and R. I. Jennrich, *The signed root deviance profile and confidence intervals in maximum likelihood analysis*, J. Am. Stat. Assoc., 91 (1996), pp. 993–998.

[5] ——, *Simple accurate approximation of likelihood profiles*, J. Comput. Graphical Statist., 11 (2002), pp. 714–732.

[6] C. de Dieuvleveult, J. Erhel, and M. Kern, *A global strategy for solving reactive transport equations*, J. Comput. Phys., 228 (2009), pp. 6395–6410.

[7] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[8] H. W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, and P. Schuster, *Inverse problems in systems biology*, Inverse Problems, 25 (2009), p. 123014.

[9] F. Fröhlich, F. J. Theis, and J. Hasenauer, *Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more.*, in Proceedings of the 12th International Conference on Computational Methods in Systems Biology (CMSB 2014), Manchester, UK, P. Mendes, J. O. Dada, and K. O. Smallbone, eds., Lecture Notes in Bioinformatics, Springer International Publishing Switzerland, Nov. 2014, pp. 61–72.

[10] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. R. Statist. Soc. B, 73 (2011), pp. 123–214.

[11] H. Haario, M. Laine, A. Mira, and E. Saksman, *DRAM: Efficient adaptive MCMC*, Stat. Comp., 16 (2006), pp. 339–354.

[12] J. Hadamard, *Sur les problèmes aux dérivées partielles et leur signification physique*, in Princeton University Bulletin, 1902, pp. 49–52.

[13] H. Hass, C. Kreutz, J. Timmer, and D. Kaschek, *Fast integration-based prediction bands for ordinary differential equation models*, Bioinf., (2016).

[14] M. Hersch, O. Hachet, S. Dalessi, P. Ullal, P. Bhatia, S. Bergmann, and S. G. Martin, *Pom 1 gradient buffering through intermolecular auto-phosphorylation*, Molecular systems biology, 7 (2015).

[15] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward, *SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers*, ACM T. Math. Software, 31 (2005), pp. 363–396.

[16] S. Hock, J. Hasenauer, and F. J. Theis, *Modeling of 2D diffusion processes based on imaging data: Parameter estimation and practical identifiability analysis*, BMC Bioinf., 14(Suppl 10) (2013).

[17] S. Hross and J. Hasenauer, *Analysis of CFSE time-series data using division-, age- and label-structured population models*, Bioinf., 32 (2016), pp. 2321–2329.

[18] S. Hug, A. Raue, J. Hasenauer, J. Bachmann, U. Klingmüller, J. Timmer, and F. J. Theis, *High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling*, Math. Biosci., 246 (2013), pp. 293–304.

[19] M. Joshi, A. Seidel-Morgenstern, and A. Kremling, *Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems*, Metabolic Eng., 8 (2006), pp. 447–455.

[20] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Applied Mathematical Sciences, Springer New York, 2006.

[21] A. Kazeroonian, F. Fröhlich, A. Raue, F. J. Theis, and J. Hasenauer, *CERENA: ChEmical REaction Network Analyzer–A toolbox for the simulation and analysis of stochastic chemical kinetics*, PLoS ONE, 11 (2016), p. e0146732.

[22] H. K. Khalil, *Nonlinear Systems*, Prentice Hall, Upper Saddle River, New Jersey, 3rd ed., 2002.

[23] C. Kreutz, A. Raue, D. Kaschek, and J. Timmer, *Profile likelihood in systems biology*, FEBS J., 280 (2013), pp. 2564–2571.

[24] R. Lockley, G. Ladds, and T. Bretschneider, *Image based validation of dynamical models for cell reorientation*, Cytometry Part A, 87 (2015), pp. 471–480.

[25] W. Q. Meeker and L. A. Escobar, *Teaching about approximate confidence regions based on maximum likelihood estimation*, Am. Stat., 49 (1995), pp. 48–53.

[26] S. A. Murphy and A. W. van der Vaart, *On profile likelihood*, J. Am. Stat. Assoc., 95 (2000), pp. 449–485.

[27] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, *Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood*, Bioinf., 25 (2009), pp. 1923–1929.

[28] A. Raue, C. Kreutz, F. J. Theis, and J. Timmer, *Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability*, Phil. Trans. Royal Soc. A, 371 (2013).

[29] F. Rigat and A. Mira, *Parallel hierarchical sampling: a general-purpose class of multiple-chains MCMC algorithms*, Comp. Stat. Data Anal., 56 (2012), pp. 1450–1467.

[30] T. E. Saunders, K. Z. Pan, A. Angel, Y. Guan, J. V. Shah, M. Howard, and F. Chang, *Noise reduction in the intracellular pom1p gradient by a dynamic clustering mechanism.*, Developmental cell, 22 (2012), pp. 558–72.

[31] J. Schöberl and W. Zulehner, *Symmetric indefinite preconditioners for saddle point problems with applications to pde-constrained optimization problems*, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 752–773.

[32] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*, Graduate studies in mathematics, American Mathematical Society, 2010.

[33] D. J. Wilkinson, *Bayesian methods in bioinformatics and computational systems biology*, Briefings in Bioinf., 8 (2007), pp. 109–116.

[34] E. Zeidler, *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*, Springer New York, 1990.