Fast derivatives of likelihood functionals for ODE based models using adjoint-state method

Reference:
Melicher Valdemar, Haber Tom, Vanroose Wim.- Fast derivatives of likelihood functionals for ODE based models using adjoint-state method
Computational statistics - ISSN 0943-4062 - 32:4(2017), p. 1621-1643
Full text (Publisher's DOI): https://doi.org/10.1007/S00180-017-0765-8

Fast derivatives of likelihood functionals for ODE based models using adjoint-state method

Valdemar Melicher*
Department of Mathematics and Computer Science, University of Antwerp
and
Tom Haber
Expertise Centre for Digital Media, Hasselt University
and
Wim Vanroose
Department of Mathematics and Computer Science, University of Antwerp

July 12, 2017

Abstract

We consider time series data modeled by ordinary differential equations (ODEs), widespread models in physics, chemistry, biology and science in general. The sensitivity analysis of such dynamical systems usually requires calculation of various derivatives with respect to the model parameters.

We employ the *adjoint state method* (ASM) for efficient computation of the first and the second derivatives of likelihood functionals constrained by ODEs. Essentially, the gradient can be computed with a cost (measured by model evaluations) that is independent of the number of the parameters and the Hessian with a linear cost in the number of the parameters instead of the quadratic one. The sensitivity analysis becomes feasible even if the parametric space is high-dimensional.

The main contributions are derivation and rigorous analysis of the ASM in the statistical context, when the discrete data are coupled with the continuous ODE model. Further, we present a highly optimized implementation of the results and its benchmarks on a number of problems.

Keywords: Sensitivity Analysis, Ordinary Differential Equations, Gradient, Hessian, Statistical Computing, Mathematical Statistics, Algorithm

*We would like to thank Xavier Woot de Trixhe from Janssen Pharmaceutica for numerous very interesting discussions on PK/PD, virology, biological pathways modeling, NLMEMs and on life in general. They were an important source of motivation and provided a view from a different perspective. The work of the first two authors was supported by IWT O&O project 130406 – ExaScience Life HPC.

1 Introduction

We consider time series vector data $\boldsymbol{y}_i \in \mathbb{R}^n$ for $i = 1, \ldots, N$, where n is the dimension of the observation space and N is the number of corresponding measurements times t_i in interval $I := [0, T]$ with some positive final time $T > 0$. Very often in science the underlying structural model for such data is the following initial-value problem

$$\begin{aligned} d_t \boldsymbol{u} &= \boldsymbol{f}(t, \boldsymbol{u}, \boldsymbol{\phi}), \quad t \in [0, T], \\ \boldsymbol{u}(0) &= \boldsymbol{u}_0(\boldsymbol{\phi}), \end{aligned} \tag{1}$$

where $\boldsymbol{u}_0$ is the initial condition, dependent only on the parameter vector $\boldsymbol{\phi} \in \mathbb{R}^p$. In general non-linear r.h.s. $\boldsymbol{f}$ of the governing equation represents the time derivative of the model variable $\boldsymbol{u}(t)$. It depends on the current time t, the model parameters $\boldsymbol{\phi}$ and the current values of $\boldsymbol{u} \in \mathbb{R}^m$.

The predictor $\hat{\boldsymbol{y}}$ of the data $\boldsymbol{y}$ is a result of integration of the dynamical system (1) and a possible subsequent post-processing, for example aggregation. This can be expressed in mathematical terms as $\hat{\boldsymbol{y}} = \mathscr{P}(\boldsymbol{u}(t, \boldsymbol{\phi})) =: \boldsymbol{g}(t, \boldsymbol{\phi})$, where $\mathscr{P} : \mathbb{R}^m \to \mathbb{R}^n$ is the post-processing operator relating the solution $\boldsymbol{u}$ to data.

The main aim of this paper is to efficiently compute the first and the second derivatives of functionals of the following form

$$l(\boldsymbol{\phi}) = \pm \sum_i d(\boldsymbol{y}_i, \boldsymbol{g}(t_i, \boldsymbol{\phi})), \tag{2}$$

with respect to $\boldsymbol{\phi}$. Here $d : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$ is a sufficiently smooth distance function (metric) on $\mathbb{R}^n$. Equation (2) measures the fidelity between the model and the data.

1.1 Motivation

The most prominent example of distance functional (2) is obtained for error model

$$\boldsymbol{y}_i = \boldsymbol{g}(t_i, \boldsymbol{\phi}) + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim_{i.i.d.} \mathscr{N}(0, \Sigma), \tag{3}$$

i.e. the residual errors $\boldsymbol{\varepsilon}_i$ are independent and identically distributed normal random variables with zero mean and residual covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. Then

$$d(\boldsymbol{y}_i, \boldsymbol{g}(t_i, \boldsymbol{\phi})) := \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{g}(t_i, \boldsymbol{\phi}))^* \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{g}(t_i, \boldsymbol{\phi})) \tag{4}$$

and we are interested in the derivatives of

$$l(\boldsymbol{\phi}) := \log p(\boldsymbol{y} | \boldsymbol{\phi}) \propto -\sum_i d(\boldsymbol{y}_i, \boldsymbol{g}(t_i, \boldsymbol{\phi})). \tag{5}$$

The gradient or the Hessian of such a log-likelihood are required quite often in statistics in various contexts. Let us supply few examples. First, Laplace's method (approximation) is very popular technique to approximate stochastic integrals of the form

$$\int e^{Ml(\boldsymbol{\phi})}\,\mathrm{d}\boldsymbol{\phi} \tag{6}$$

around a mode $\hat{\boldsymbol{\phi}}$ of sufficiently smooth negative function l leading to

$$\int e^{Ml(\boldsymbol{\phi})}\,\mathrm{d}\boldsymbol{\phi} \to \left(\frac{2\pi}{M}\right)^{p/2} |-H(\hat{\boldsymbol{\phi}})|^{-1/2} e^{Ml(\hat{\boldsymbol{\phi}})} \tag{7}$$

as $M(\in \mathbb{R}^+) \to \infty$ (Wong, 2001). The evaluation of Hessian $H(\hat{\boldsymbol{\phi}})$ of l is needed. Second, when looking for a maximum-likelihood estimator $\hat{\boldsymbol{\phi}}^{MLE} := \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\, l(\boldsymbol{\phi})$ one usually applies some optimization algorithm which requires many evaluations of gradient $\nabla l(\boldsymbol{\phi})$, such as CG or BFGS (Bertsekas, 1999; Bazaraa et al., 2006).

Third, modern Monte Carlo Markov Chain (MCMC) samplers such as Metropolis-Adjusted Langevin algorithm (MALA) or Hamiltonian Monte Carlo (HMC) again require computing gradients or even Hessians of a log-likelihood with respect to the model parameters for every sample (Brooks et al., 2011).

For any of the above problems, the computation of the derivatives is a key operation and its speedup directly translates to the speedup of the whole algorithm. For example in the case of the HMC sampler, the total speedup is roughly proportional to the speedup of the gradient computation of the log-likelihood, one can achieve.

1.2 Adjoint-state method

To efficiently compute the first and the second derivative (gradient and Hessian) of (2) with respect to the model parameters $\boldsymbol{\phi}$ we will employ *adjoint-state method* (ASM).

The ASM is used in many different fields, such as control theory (Lions, 1971), data assimilation in meteorology (Lewis et al., 2006) or parameter identification (Melicher and Vrábel', 2013; Cimrák and Melicher, 2007). It is difficult to precisely trace its origin, since it is based on a general principle - the duality. Special dual problems or special test functions in general are used extensively in functional analysis for quite different tasks, e.g. in homogenization theory (Bensoussan et al., 1978). The idea of the ASM is to derive a special dual problem to the sensitivity equation of (1), which allows one to write the derivative(s) of (2) in a simple form which is inexpensive to evaluate. Usually, one obtains an inner product(s) in a suitable Hilbert space containing the dual state.

Even if the ASM is a classical method in many different fields, its applications in general statistical literature are rather scarce or could be even considered virtually non-existent. In our opinion, this is due to several reasons.

Mainly, it is a matter of need. Until recently the usual statistical models had only few parameters and the corresponding derivatives were easily evaluable using finite differences. On the other hand, the ASM was mainly used for problems, where derivatives with respect to infinite dimensional parameters are needed, such as the optimal control of partial differential equations (PDEs) (Lions, 1971). For those problems, the ASM is the only viable way to compute the gradient of a cost functional.

The second possible reason is the lack of interdisciplinary publications in statistical literature with the fields where the ASM plays the role of a classical technique. One of the exceptions that elegantly connects the worlds of the PDE-constrained inverse problems and that of Bayesian inference is the paper (Martin et al., 2012). Meteorology is probably the field where the relationship between differential models and stochastic processes is the most advanced (Lewis et al., 2006).

The third and probably rather influential reason is that the results presented in literature regarding the ASM do not take into account the specifics of statistical estimation, particularly that the measurements can not be altered or interpolated in any way. In this paper, we present an ASM framework for ODE based statistical models, which recognizes and resolves this issue. The ODE case can be addressed in generality, which is not possible for PDE-constrained problems.

The dynamical models described by ODEs are rather widely used in science. They are simply indispensable for acquiring essential knowledge about complex biological systems (Murray, 2002; Draelants et al., 2012) as is the case for other fields studying intricate matters such as psychology and economics. In chemistry, regardless of the criticism (Gillespie, 1977), the reaction-rate equations [1] are still extensively employed. We are motivated by applications in PK/PD modeling and virology (Lavielle et al., 2011; Tornøe et al., 2004).

The sensitivity analysis of ODEs is well established in literature. Let us only mention a classical book on the optimal control of ODEs (Cesari, 1983). Moreover, many results that are intended for PDEs are directly applicable to ODEs, since from the mathematical point of view, an ODE could be simply seen as a PDE without a spatial differential operator. However, as already mentioned, the relevant results presented in literature, do not take into account the specifics of statistical estimation.

The ASM is usually applied in a PDE-constrained context. The fidelity between the data and the PDE-based model is measured in a Lebesgue space L^p-norm, particularly in L^2 sense, as is also the case of the above mentioned paper (Martin et al., 2012). It implies that the data are considered to be defined almost everywhere in the space or in the space-time in the case of time-dependent problems. This is however in a strong contrast with statistical philosophy. The measurements are ultimately discrete and sacred. E.g. by interpolating the measurements, new ones are generated and that can not be tolerated.

The main contribution of this paper is that it recognizes and resolves this problem. We show, that the discrete data $\boldsymbol{y}$ can be combined with the continuous model (1) at the level of the likelihood functional (2).

[1] Considering spatial phenomena such as diffusion and(or) convection leads to PDE models, see for example (Slodička and Balážová, 2010).

The resulting adjoint problem contains a Dirac delta source corresponding to individual measurement times. The developments are fully supported by rigorous proofs.

The subsequent numerical analysis shows that the ASM application for statistical estimation is far from obvious and more work is still needed to reach its efficiency in the deterministic setting.

Moreover, only the first order results are normally available, since the ASM is usually applied in infinite dimensional setting, as explained above. We supply the Hessian computation as well.

Another contribution is a highly efficient implementation of the results and its benchmarking with respect to finite differences and sensitivity equation approach.

To our best knowledge, we do not know about similar results in the literature.

Last but not least, this interdisciplinary paper aims to popularize this quite underused but potentially very useful method in the statistical community and help those working with ODE based models to compute the corresponding ODE-model sensitivities more efficiently. Recently, with boom in general availability and dimensionality of data, ODE models with a high number of parameters are being employed. The evaluation of gradients becomes very costly and consequently various derivative-free methods have become more popular, see for example Delyon et al. (1999). The application of the ASM makes derivative-based algorithms again competitive. The gradient can be computed with a cost (measured by model evaluations) that is essentially independent of the number of the parameters. The Hessian can be computed as well, with essentially linear cost in the number of the parameters instead of the quadratic one.

The paper is divided as follows. In Section 2 we analyze the sensitivity of initial value problem (1) with respect to its parameter vector $\boldsymbol{\phi}$. In Section 3, we present in detail the here mentioned approach to combine discrete data with continuous model. Then in Section 4 we obtain the ASM for computing of the gradient and Hessian of (2) with respect to $\boldsymbol{\phi}$. In Section 5 its implementation is discussed and in Section 6 its efficiency is tested on a number of examples.

2 Sensitivity of model

In this preparatory section we will discuss the well-posedness of the initial value problem (1) as well as the existence of its derivative with respect to the parameters $\boldsymbol{\phi}$. We follow the presentation in (Zeidler, 1985) with all the relevant notation, so we can be rather concise.

The first Gatèaux differential of a function f with respect to $\boldsymbol{x}$ in direction $\boldsymbol{h}$ is denoted by $\mathscr{D}f(\boldsymbol{x};\boldsymbol{h})$. Then, let us denote by $\boldsymbol{s} := \mathscr{D}\boldsymbol{u}(\boldsymbol{\phi};\boldsymbol{h})$, i.e. the first Gatèaux differential (we will show it is Fréchet as well) of the model function $\boldsymbol{u}$ with respect to the parameters $\boldsymbol{\phi}$ in direction $\boldsymbol{h}$. If it exists, the formal differentiation of (1) yields that $\boldsymbol{s}$ is the solution to the following initial value problem

$$\begin{aligned} d_t\boldsymbol{s} &= J_{\boldsymbol{u}}(\boldsymbol{f})\boldsymbol{s} + J_{\boldsymbol{\phi}}(\boldsymbol{f})\boldsymbol{h}, \quad t \in [0,T], \\ \boldsymbol{s}(0) &= J_{\boldsymbol{\phi}}(\boldsymbol{u}_0)\boldsymbol{h}, \end{aligned} \tag{8}$$

known as the *sensitivity equation*. Here $J_{\boldsymbol{\phi}}(\boldsymbol{f}) : \mathbb{R}^p \to \mathbb{R}^m$ and $J_{\boldsymbol{u}}(\boldsymbol{f}) : \mathbb{R}^m \to \mathbb{R}^m$ are the Jacobians of the r.h.s. $\boldsymbol{f}$ of the model (1) with respect to the model parameters $\boldsymbol{\phi}$ and the state variables of the model $\boldsymbol{u}$, respectively. The analogical is the meaning of $J_{\boldsymbol{\phi}}(\boldsymbol{u}_0)$.

Let $\boldsymbol{e}_i, i = 1,\dots,p$ be the canonical basis in $\mathbb{R}^p$. Solving (8) for $\boldsymbol{h} = \boldsymbol{e}_i$ for each $i = 1,\dots,p$ yields $\boldsymbol{s} = (\frac{\partial \boldsymbol{u}_1}{\partial \boldsymbol{\phi}_i}, \frac{\partial \boldsymbol{u}_2}{\partial \boldsymbol{\phi}_i}, \dots, \frac{\partial \boldsymbol{u}_m}{\partial \boldsymbol{\phi}_i})^*$, if the partial derivatives exist. It means that to compute the whole jacobian $J_{\boldsymbol{\phi}}(\boldsymbol{u})$ one needs to integrate p initial value problems (8). The complexity is essentially identical to that of the first-order finite difference approximation, as will be confirmed in Section 6. The sensitivity equation approach is however still preferred if high accuracy is needed.

Let us restate the Theorem 4.D from (Zeidler, 1985) in our context.

Theorem 1. *Suppose that the mappings $\boldsymbol{f} : U \subseteq \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^m$ and $\boldsymbol{u}_0 : V \subseteq \mathbb{R}^p \to \mathbb{R}^m$ are $C^k, k \geq 1$ and that U, V are open sets containing $(0, \boldsymbol{u}_0(\boldsymbol{\phi}_0), \boldsymbol{\phi}_0)$ and $\boldsymbol{\phi}_0$, respectively. Then:*

- *(a) There exists an interval $(-a,a), a > 0$, and an open neighborhood $U(\boldsymbol{\phi}_0)$ such that the initial value problem (1) has exactly one solution for each $\boldsymbol{\phi} \in U(\boldsymbol{\phi}_0)$.*
- *(b) The mapping $(t,\boldsymbol{\phi}) \mapsto \boldsymbol{u}(t;\boldsymbol{\phi})$ is C^k on $(-a,a) \times U(\boldsymbol{\phi}_0)$, and (8) holds.*

Since our initial value problem (1) slightly differs from that of (Zeidler, 1985) and also for the completeness we present a proof in Appendix A.

From now on, any formal differentiation of $\boldsymbol{u}$ with respect to the parameters $\boldsymbol{\phi}$ is justified by Theorem 1. The theorem provides only a local result regarding the existence and the uniqueness of the solution $\boldsymbol{u}$ of the ivp (1). Consequently, we have to assume that $T < a$.

3 Connecting the worlds

Measurements $\boldsymbol{y}_i$ are acquired at discrete time points t_i. In statistics, these measurements should not be tempered with in any way, e.g. they cannot be interpolated, which stands for augmentation.

On the other hand the model (1) is a continuous one and since the adjoint-state method (ASM) deals extensively with the model and the functional (2), it is necessary to work in continuous setting.

We will connect the discrete data and the continuous model on the level of the likelihood functional. One can write

$$\sum_i d(\boldsymbol{y}_i, \boldsymbol{g}(t_i,\boldsymbol{\phi})) = \int_0^T \delta\{t - t_i\} d(\boldsymbol{y}(t), \boldsymbol{g}(t,\boldsymbol{\phi}))\,\mathrm{d}t \tag{9}$$

where, by the classical misuse of notation, $\delta\{t - t_i\}$ is the Dirac delta function of the set of all measurement times t_i. In order to achieve that the above integral is well-defined, we will consider a small positive ε, such that the functions $\boldsymbol{y}(t) := \boldsymbol{y}_i$, $t \in (t_i - \varepsilon, t_i + \varepsilon)$ for all measurement times t_i are well defined. We emphasize, that by doing so, we do not generate new measurements. We merely assume an infinitesimally small interval of their validity. The $\boldsymbol{y}(t)$-values outside of intervals $t \in (t_i - \varepsilon, t_i + \varepsilon)$ are irrelevant.

For clarity, we extend the function $\boldsymbol{y}(t)$ outside of these intervals by linear interpolation to continuous functions on whole interval $[0,T]$ [2].

For the well-posedness of (9), also the model $\boldsymbol{g}(t,\boldsymbol{\phi})$ has to be at least continuous around each t_i. Let us assume that $\mathscr{P} \in \mathscr{L}(\mathbb{R}^m,\mathbb{R}^n)$, i.e. $\mathscr{P}$ is a linear operator from $\mathbb{R}^m$ to $\mathbb{R}^n$. The linearity is a sufficient condition for the validity of

$$\mathscr{P}(\boldsymbol{u}(t,\boldsymbol{\phi})) - \mathscr{P}(\boldsymbol{u}_0(\boldsymbol{\phi})) = \int_0^t \mathscr{P}(\boldsymbol{f}(t,\boldsymbol{\phi},\boldsymbol{u}))\,\mathrm{d}t \tag{10}$$

which will be needed for the subsequent developments. From now on we write $\mathscr{P}\boldsymbol{u}$ instead of $\mathscr{P}(\boldsymbol{u})$. Let us point out that this assumption is usually not restrictive in practical applications. The eventual non-linear transformations can be applied a priori to the data $\boldsymbol{y}$ or included in $\boldsymbol{f}$. Now, since the solution $\boldsymbol{u}$ to (1) is at least continuously differentiable (Theorem 1) and a linear operator preserves continuity, $\boldsymbol{g}(t,\boldsymbol{\phi})$ is trivially continuous.

As a convenience, for any distribution d and sufficiently smooth function f, we denote by $\langle d, f\rangle$ the duality between them on the time interal $[0,T]$. We will also need the scalar product $(\cdot,\cdot)$ in the Hilbert space $L^2([0,T])$. Using this notation, (9) can be rewritten as

$$\int_0^T \delta\{t-t_i\} d(\boldsymbol{y}(t),\boldsymbol{g}(t,\boldsymbol{\phi}))\,\mathrm{d}t = \langle \delta\{t-t_i\}, d(\boldsymbol{y}(t),\boldsymbol{g}(t,\boldsymbol{\phi}))\rangle\,. \tag{11}$$

Let us introduce short notation $\{\delta\}$ for $\delta\{t-t_i\}$. At last, the equality (9) defines a seminorm on $C([0,T],\mathbb{R}^n)$, since the l.h.s. is a discrete norm. We denote this seminorm simply as $\|\cdot\|$.

Using the gluing notation above, we can prove the following lemma that allows us to evaluate the first differential of (2) using the solution $\boldsymbol{s}$ to the sensitivity equation (8).

Lemma 1. *Let the assumptions of Theorem 1 be fulfilled for $k=1$ and let the metric d be C^1. Then the functional (2) is Fréchet differentiable and the differential $\mathscr{D}l(\boldsymbol{\phi};\boldsymbol{h})$ can be expressed as*

$$\mathscr{D}l(\boldsymbol{\phi};\boldsymbol{h}) = \langle \{\delta\} d_{\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{g}(t,\boldsymbol{\phi})),\boldsymbol{s}\rangle_{,} \tag{12}$$

where $\boldsymbol{s}$ is the unique solution to sensitivity equation (8).

A proof can be found in Appendix A. Moreover, due to the linearity of (8), the differential $\mathscr{D}l(\boldsymbol{\phi};\boldsymbol{h})$ can be easily written in the linearized form $\mathscr{D}l(\boldsymbol{\phi};\boldsymbol{h}) = l'(\boldsymbol{\phi})\boldsymbol{h}$. Since we work in finite dimensional spaces, the expression

$$\nabla l(\boldsymbol{\phi})\cdot\boldsymbol{h} := l'(\boldsymbol{\phi})\boldsymbol{h} \quad \text{for all } \boldsymbol{h}\in\mathbb{R}^p \tag{13}$$

well defines the gradient $\nabla l(\boldsymbol{\phi})$ of l as an element in $\mathbb{R}^p$ for each fixed $\boldsymbol{\phi}$, i.e. $\nabla l : \mathbb{R}^p \to \mathbb{R}^p$.

[2] Other continuous "interpolation" are possible such as piecewise-linear or by cubic splines, but they are less graphic.

As explained in Section 2, p initial value problems (8) have to be computed to evaluate $\nabla l(\boldsymbol{\phi})$ for some $\boldsymbol{\phi}$.

4 Adjoint-state method

In this Section we will introduce *adjoint-state method* (ASM) for computation of the gradient and the Hessian of (2). The results are strongly influenced by the peculiar coupling between the discrete measurements and the continuous model (1). Let us directly present the main statement.

Theorem 2. *Let the assumptions of Lemma 1 be fulfilled. Then the first Fréchet differential in (12) can be also written as*

$$\mathscr{D}l(\boldsymbol{\phi};\boldsymbol{h}) = -\boldsymbol{v}^*(0)J_{\boldsymbol{\phi}}(\boldsymbol{u}_0)\boldsymbol{h} - \left(J_{\boldsymbol{\phi}}(\boldsymbol{f})\boldsymbol{h},\boldsymbol{v}\right) \tag{14}$$

where $\boldsymbol{v}$ is the unique solution to the following initial value problem

$$\begin{aligned} d_t\boldsymbol{v} &= -J^*_{\boldsymbol{u}}(\boldsymbol{f})\boldsymbol{v} + \{\delta\}d_{\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{g}(t,\boldsymbol{\phi})), \quad t\in[0,T], \\ \boldsymbol{v}(T) &= 0. \end{aligned} \tag{15}$$

A proof is again presented in Appendix A. Obviously, Equation (14) is written in linearized form. We get

$$\nabla l = -\boldsymbol{v}^*(0)J_{\boldsymbol{\phi}}(\boldsymbol{u}_0) - \left(\boldsymbol{v},J_{\boldsymbol{\phi}}(\boldsymbol{f})\right), \tag{16}$$

where the second term on the r.h.s. is a vectorial integral. This is a very efficient way to compute the gradient. One has to only integrate one adjoint problem (15) and evaluate the expression (16), i.e. to compute p scalar products in $L^2([0,T])$.

Let us discuss the result a little. The ivp (15) is a special ODE. First, it has absolutely no physical, chemical, biological or any other interpretation of the underlying scientific field of equation (1). The best way to look at it is that it is a special dual problem (see the proof) to the sensitivity equation (8), which allows us to efficiently compute the gradient of (2) (and the Hessian as well as we will see.) Then, it is a final time problem to be integrated from T to the initial time 0. It is a linear ODE like the sensitivity equation. Its r.h.s. contains the term $\{\delta\}d_{\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{g}(t,\boldsymbol{\phi}))$, which expresses how quickly the distance between the data and the model changes when changing the model variable $\boldsymbol{u}$.

Probably the most important fact to note about the ASM is that it operates at a higher level than the sensitivity equation method. It does not supply the derivative of the state $\boldsymbol{u}$, but directly the one of the likelihood functional (2). By considering the model together with (2), the efficiency can be achieved.

The numerical issues regarding the integration of (15) will be discussed in Section 5.

Example 1. *Let us consider the distance (4) corresponding to the multivariate normal distribution of* *the residual errors. Then the derivative $d_{\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{g}(t,\boldsymbol{\phi}))$ in the r.h.s of (15) reads*

$$d_{\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{g}(t,\boldsymbol{\phi})) = -\mathscr{P}^{*}\Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{g}(t_i,\boldsymbol{\phi})). \tag{17}$$

The adjoint problem is dependent on the residual covariance matrix Σ and on the post-processing operator $\mathscr{P}$.

4.1 Evaluating Hessian

Let us depict the second Gatèaux differential of a functional f with respect to $\boldsymbol{x}$ in directions $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ as $\mathscr{D}^2 f(\boldsymbol{x};\boldsymbol{h}_1,\boldsymbol{h}_2)$. Then, let us introduce notation $\boldsymbol{\varsigma} := \mathscr{D}^2\boldsymbol{u}(\boldsymbol{\phi};\boldsymbol{h}_1,\boldsymbol{h}_2)$. We will show that $\boldsymbol{\varsigma}$ is Fréchet as well. By formally differentiating (8) one more time with respect to $\boldsymbol{\phi}$ we obtain that $\boldsymbol{\varsigma}$ is a solution to the following initial value problem

$$\begin{aligned} d_t\boldsymbol{\varsigma} &= \boldsymbol{f}_{\boldsymbol{\phi}\boldsymbol{\phi}}\boldsymbol{h}_1\boldsymbol{h}_2 + \boldsymbol{f}_{\boldsymbol{\phi}\boldsymbol{u}}\boldsymbol{h}_1\boldsymbol{s}_2 + \boldsymbol{f}_{\boldsymbol{u}\boldsymbol{\phi}}\boldsymbol{s}_1\boldsymbol{h}_2 \\ &\quad + \boldsymbol{f}_{\boldsymbol{u}\boldsymbol{u}}\boldsymbol{s}_1\boldsymbol{s}_2 + J_{\boldsymbol{u}}(\boldsymbol{f})\boldsymbol{\varsigma}, \quad t\in[0,T], \\ \boldsymbol{\varsigma}(0) &= (\boldsymbol{u}_0)_{\boldsymbol{\phi}\boldsymbol{\phi}}\boldsymbol{h}_1\boldsymbol{h}_2 \end{aligned} \tag{18}$$

known as the *second sensitivity equation*. Here $\boldsymbol{s}_1$, $\boldsymbol{s}_2$ are the solutions to (8) for $\boldsymbol{h}=\boldsymbol{h}_1$, $\boldsymbol{h}=\boldsymbol{h}_2$, respectively. The second order derivatives in (18) are essentially three-dimensional tensors. In Appendix A the following lemma is proven.

Lemma 2. *Let the assumptions of Theorem 1 be fulfilled for $k=2$ and let the metric d be C^2. Then the second Fréchet differential of (2) with respect to the model parameters $\boldsymbol{\phi}$ can be written as*

$$\mathscr{D}^2 l(\boldsymbol{\phi};\boldsymbol{h}_1,\boldsymbol{h}_2) = \left\langle\{\delta\}, d^2_{\boldsymbol{u}\boldsymbol{u}}\boldsymbol{s}_1\boldsymbol{s}_2\right\rangle + \left\langle\boldsymbol{\varsigma},\{\delta\}d_{\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{g}(t,\boldsymbol{\phi}))\right\rangle, \tag{19}$$

where $\boldsymbol{\varsigma}$ is the unique solution to (18). Moreover, the second term in (19) can be rewritten using the solution $\boldsymbol{v}$ to (15) as

$$\begin{aligned} \left\langle\boldsymbol{\varsigma},\{\delta\}d_{\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{g}(t,\boldsymbol{\phi}))\right\rangle &= -(\boldsymbol{u}_0)_{\boldsymbol{\phi}\boldsymbol{\phi}}\boldsymbol{h}_1\boldsymbol{h}_2\cdot\boldsymbol{v}(0) - \left(\boldsymbol{f}_{\boldsymbol{\phi}\boldsymbol{\phi}}\boldsymbol{h}_1\boldsymbol{h}_2,\boldsymbol{v}\right) \\ &\quad - \left(\boldsymbol{f}_{\boldsymbol{\phi}\boldsymbol{u}}\boldsymbol{h}_1\boldsymbol{s}_2,\boldsymbol{v}\right) - \left(\boldsymbol{f}_{\boldsymbol{u}\boldsymbol{\phi}}\boldsymbol{s}_1\boldsymbol{h}_2,\boldsymbol{v}\right) - \left(\boldsymbol{f}_{\boldsymbol{u}\boldsymbol{u}}\boldsymbol{s}_1\boldsymbol{s}_2,\boldsymbol{v}\right). \end{aligned} \tag{20}$$

Solving (18) for $\boldsymbol{h}_1=\boldsymbol{e}_i$, $\boldsymbol{h}_2=\boldsymbol{e}_j$ for each $i=1,\ldots,p$, $i=1,\ldots,p$ yields $\boldsymbol{\varsigma}=(\frac{\partial^2\boldsymbol{u}_1}{\partial\boldsymbol{\phi}_i\partial\boldsymbol{\phi}_j},\frac{\partial^2\boldsymbol{u}_2}{\partial\boldsymbol{\phi}_i\partial\boldsymbol{\phi}_j},\ldots,\frac{\partial^2\boldsymbol{u}_m}{\partial\boldsymbol{\phi}_i\partial\boldsymbol{\phi}_j})^{*}$. It means that to compute the Hessian $H_{\boldsymbol{\phi}}(\boldsymbol{u})$, one needs to integrate the second sensitivity equation (18) $p(p+1)/2$ times. For that one moreover needs to compute p sensitivities $\boldsymbol{s}_i$ for each $h=\boldsymbol{e}_i$, $i=1,\ldots,p$. The cost is essentially identical to that of the first order finite difference approximation. Again, it is beneficial if high accuracy is needed.

On the other hand, the evaluation of the Hessian $H_{\boldsymbol{\phi}} l$ of (2) via (20) requires us to only compute one adjoint problem (15), p sensitivity equations (8) and $p(p+1)/2$ times the four scalar products from (20). This is a very efficient and accurate way how to compute the Hessian.

As before with the gradient, we see that the ASM supplies the sensitivity at the level of the functional, not that of the underling model state $\boldsymbol{u}$.

4.1.1 Hessian via adjoint with finite differences

Let us present an alternative way to efficiently compute the Hessian of (2), which is slightly less accurate than using (20) but much simpler to implement. The idea is to combine (14) with finite differences as follows

$$H_i(l(\boldsymbol{\phi})) \approx \frac{\nabla l(\boldsymbol{\phi} + h\boldsymbol{e}_i) - \nabla l(\boldsymbol{\phi})}{h}, \tag{21}$$

where H_i stands for the i-th column of H (or row) and h is a small positive real number. We recall that $\{\boldsymbol{e}_i : 1 \leq i \leq p\}$ is the canonical basis in $\mathbb{R}^p$. Each of p gradients $\nabla l(\boldsymbol{\phi} + h\boldsymbol{e}_i)$, $1 \leq i \leq p$ is computed using (14). Together $p+1$ adjoint initial value problems (15) need to be integrated.

5 Implementation of ASM

In this Section we will describe an implementation of the ASM presented in Section 4. At the core of the developments is the adjoint initial value problem (15). Even if it is a rather simple linear ODE-system, it is a quite difficult one to solve numerically because of its r.h.s. containing the Dirac delta function source term.

Our implementation closely follows the constructive proof of Theorem 2 in Appendix A. At each measurement time t_i, ODE-solver is stopped, $d_{\boldsymbol{u}}(\boldsymbol{y}_i, \boldsymbol{g}(t_i, \boldsymbol{\phi}))$ is explicitly added to the current solution and then the integration is resumed. We solve a sequence of initial value problems (32) instead of the original ivp (15).

Unfortunately, the repetitive restarting of the ODE solver has a negative impact on the performance. The "energy of the measurement residual" is added to the dynamical system at once via the initial condition. The time derivative in (32) is proportional to $\boldsymbol{v}(t_i)$. The steepness of the solution forces the ODE-solver to advance in many small time steps, which leads to a high number of iterations. We will see in Section 6, that the efficiency of ASM is indeed strongly dependent on the number of measurements.

However, equation (32) is a quite simple linear ODE-system, which should be exploitable in multiple ways. Though the increasing of the numerical efficiency of backward integration while preserving the statistical rigor is a very interesting scientific goal, it is out of the scope of this contribution and left for the future research.

We tackle (15) using CVODES solver from the SUNDIALS package (Hindmarsh et al., 2005). The leading author of the development team, Alan Hindmarsh, is the creator of the famous LSODE - Livermore Solver for Ordinary Differential Equations (Hindmarsh, 1980). CVODE and CVODES are written in ANSI standard C (LSODE is in Fortran 77). CVODES is an extended CVODE code with both forward and backward sensitivity abilities (Serban and Hindmarsh, 2005).

Remark 1. *Let us imagine, we would not explicitly integrate the Dirac delta function out. It can be approximated in many different ways, but the most suitable (uninformative) from the statistical point of view is the approximation using Gaussian*

$$\delta_i^\sigma(t) := \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-t_i)^2}{2\sigma^2}}, \tag{22}$$

where σ can be seen as a measure of the trust that the measurements have been taken precisely at the times t_i. Let us define $\delta^\sigma(t) := \sum_i \delta_i^\sigma(t)$. Using this approximation, the adjoint system (15) becomes

$$\begin{aligned} d_t \boldsymbol{v} &= -J_{\boldsymbol{u}}^*(\boldsymbol{f})\boldsymbol{v} + \delta^\sigma(t) d_{\boldsymbol{u}}(\boldsymbol{y}, \boldsymbol{g}(t, \boldsymbol{\phi})), \quad t \in [0, T], \\ \boldsymbol{v}(T) &= 0. \end{aligned} \tag{23}$$

Here, the adjoint problem (23) represents an interesting antagonism between the efficiency of the ASM and the statistical rigorousness one expects. The higher the trust in the measurement times, the smaller the σ and consequently higher the derivative of the r.h.s of (23) with respect to time which makes this dynamical system more and more difficult for an ODE solver to integrate.[3]

6 Numerical experiments

In this section we will consider for simplicity but without any loss of generality the Gaussian log-likelihood (5) with $\Sigma = I$.

All the numerical experiments presented in this Section are computed in $Diff$MEM. It is a package for the fitting of mixed-effect models constrained by differential equations. The package is under active development by the authors and the algorithms presented in this paper are only a subset of its abilities.

At present, only ODE dynamical models are supported. $Diff$MEM employs CVODE ODE solvers for quick and robust solution of those models. It uses Eigen linear algebra package (library) to represent its internal memory containers and to solve underlying linear systems.

For all the experiments in this Section, the ODE solvers' absolute accuracies are set to 10^{-14} and the relative ones to 10^{-10}. These accuracies are sufficient to remove considerations about accuracy of the ODE-solver from the analysis.

We will study the efficiency and robustness of adjoint-state method (ASM) for computing the derivatives of the likelihood.

[3] The variance σ^2 has here a purely ad hoc use for the above argument. We are not interested if it is prescribed or estimated from the data.

6.1 Linear model

To start, let us consider the classical linear ordinary differential equation (ODE)

$$\begin{aligned} d_t\boldsymbol{u} &= A\boldsymbol{u}, \quad t \in [0,T], \\ \boldsymbol{u}(0) &= \boldsymbol{u}_0, \end{aligned} \tag{24}$$

where A is a $d \times d-$dimensional matrix, the elements of which represent the model parameters. This simple model is ideal toy-example to comprehend the importance of ASM for models with high dimensional parametric space.

Let us consider diagonal matrix A. Then the dimensions of the parametric space and of the solution coincide ($p = m$). Moreover, we can easily calculate the exact solution

$$\boldsymbol{u}_i = \boldsymbol{u}_{0,i} e^{\boldsymbol{\phi}_i t}, \quad i = 1, \ldots, p, \tag{25}$$

where $\boldsymbol{\phi}_i = A_{ii}$.

We consider 13 different dimensions of $\boldsymbol{\phi}$, ranging from 2 to 122. For each of them we have randomly generated 100 parameter-samples $\boldsymbol{\phi}$ as follows

$$\boldsymbol{\phi}_i \sim U[-1.1, -0.1], \quad 1 \leq i \leq p. \tag{26}$$

We set $\boldsymbol{u}_0 = \mathbf{1}$. The number of observation time points N, regularly spread in $[0,100]$, is set constant to 11.

The corresponding synthetic data $\boldsymbol{y}$ are also perturbed as follows:

$$\boldsymbol{y}_i \sim U[\boldsymbol{y}_i, \boldsymbol{y}_i + 10^{-1}\max(\boldsymbol{y})], \quad 1 \leq i \leq N, \tag{27}$$

where $\max(\boldsymbol{y}) \in \mathbb{R}^n$ is a constant vector containing at each position the same maximum. We would like to emphasize that any reasonable perturbation leads to the same results. It is only important that the data are perturbed outside of the log-likelihood mode.

We have computed the gradient of likelihood using finite differences (FD), the ASM approach (14) and using the sensitivity equation (SE) (8). We recall that the last two approaches are implemented using the CVODES forward- and backward- sensitivity abilities and all the thinkable settings are identical to make comparison as sound as possible. The results are presented in Figure 1.

The timings of FD, ASM and of SE are presented in Figure 1(a). The first conclusion is that our implementation of SE-approach is optimal since the timings of FD and SE more or less coincide. Actually SE is always a slightly faster method. Given the significantly higher accuracy of SE with respect to FD (1(d)), it renders FD-approach redundant.

Somewhere around 10 parameters ASM becomes on average more efficient than SE. Moreover, given

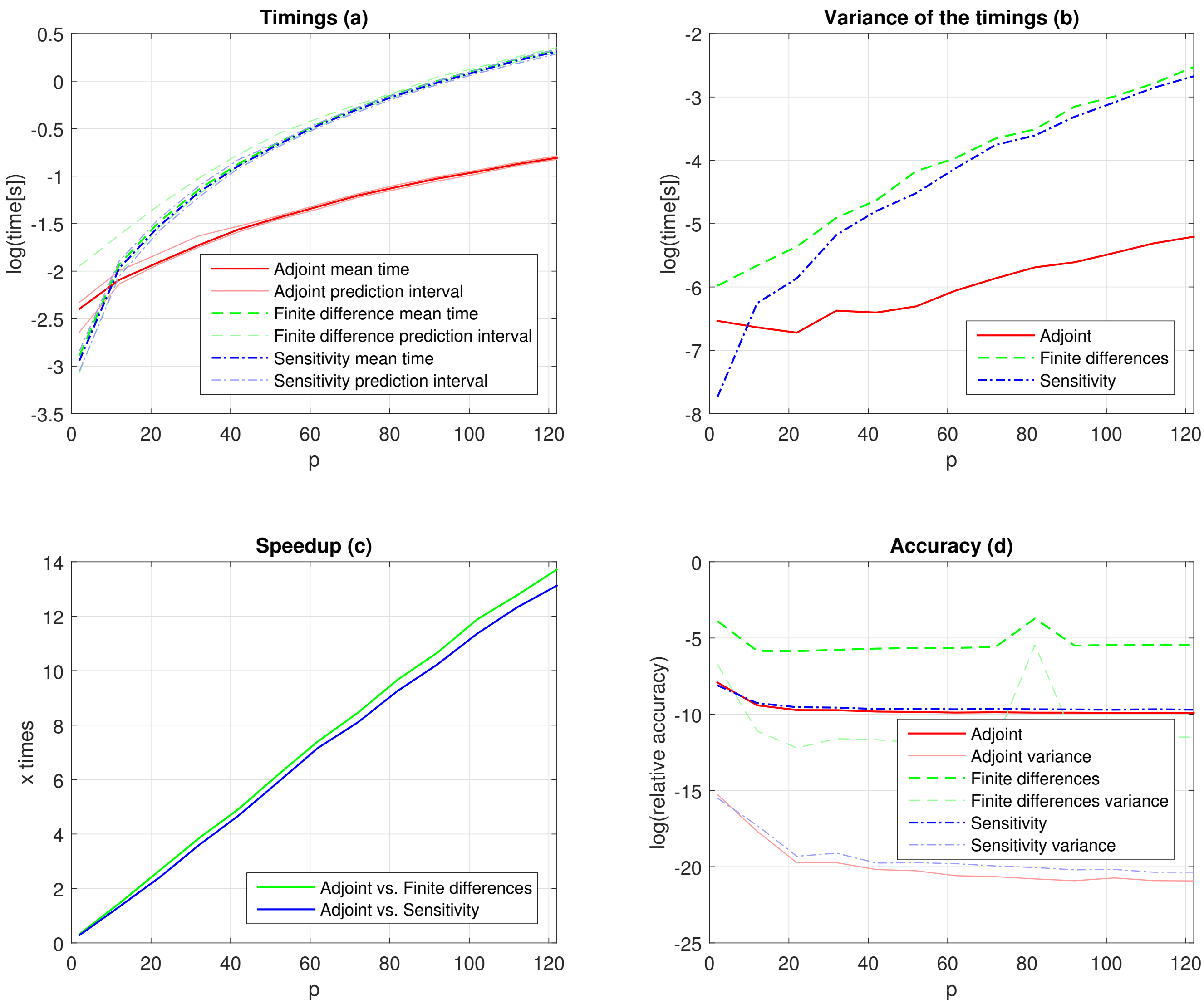

Figure 1: Adjoint gradient for linear ODE model: parameter space dimension

the non-parametric prediction intervals based on the 100 samples, it is from around 15 parameters virtually always more time-efficient than SE. This reasoning is conservative since it does not take the correlation between ASM and SE timings into account. Moreover, the variance of timings is for $\dim(\boldsymbol{\phi}) > 10$ lower for ASM than for SE, see Figure 1(b), rendering timing predictions for ASM more reliable.

The time efficiency of both ASM and SE is negatively impacted by exclusive use of dense matrices in $Diff$MEM. The equations (8), (15), (14) require evaluation of Jacobians $J_{\boldsymbol{u}}(\boldsymbol{f})$ and $J_{\boldsymbol{\phi}}(\boldsymbol{f})$. These are extremely sparse [4]. More importantly, because of the dense matrix implementation only rather small systems can be solved. Sparse matrix implementation is planned for the future versions of $Diff$MEM. Both ASM and SE are influenced to the same degree and the relative comparison holds.

The speedup of ASM vs. SE (1(c)) is roughly linear in the number of the parameters but it slows down slightly for higher parameter dimensions. The suspected cause here is the cost of memory access when CVODES evaluates the forward solution $\boldsymbol{u}$ during the backwards integration of (15).

The accuracy of both ASM and SE with respect to the exact gradient of the likelihood (5) is presented in Figure 1(d). Both methods achieve virtually identical accuracy since the forward- and backward-solvers use the same relative and absolute tolerances.

Now we will examine the efficiency of ASM and SE with respect to the number of time observations. We fix the dimension of the problem at e.g. $\dim(\boldsymbol{\phi}) = 50$. The number of time observations $\dim(y)$ in $[0, 100]$ fluctuates between 2 and 122 in 12 steps. For each number we again compute 100 gradients using (8) and (14). The parameters $\boldsymbol{\phi}$ are again generated using (26). The results are presented in Figure 2.

The SE approach efficiency is essentially independent of the number of time observations (2(a).) The ASM efficiency however decreases with increasing number of observations. As explained in Section 5, the backward adjoint integrator needs after each data point a large number of small time steps to account for the steepness of the adjoint solution $\boldsymbol{v}$. The negative impact is the most clearly visible in Figure 2(c). For many practically relevant problems [5], the number of measurements is rather low, making this issue less pronounced. Anyhow, increasing the numerical efficiency of backward integration while preserving the statistical rigor is obviously a very interesting direction for future research.

Implicitly, since for the diagonal linear model $p = n$, also the dimension of the solution space plays a role. But we do not compare the speed and accuracy of the different methods with respect to m or n, since their complexities with respect to these are the same.

Now, again using the problem (24), we will illustrate the efficiency of computing the Hessian of (5) with respect to $\boldsymbol{\phi}$ employing the expression (20). We compare this (SA approach) with Hessian evaluated using the finite difference approximation (FD) and the one computed using (21) (FA).

Remark 2. *First-order Gauss-Newton approximation of the Hessian, where the second term in (19) is neglected, is not included in the comparison. When the model does not yet well approximate the data,*

[4] The diagonal system (24) is the most sparse system one can think of.

[5] PK/PD, virology.

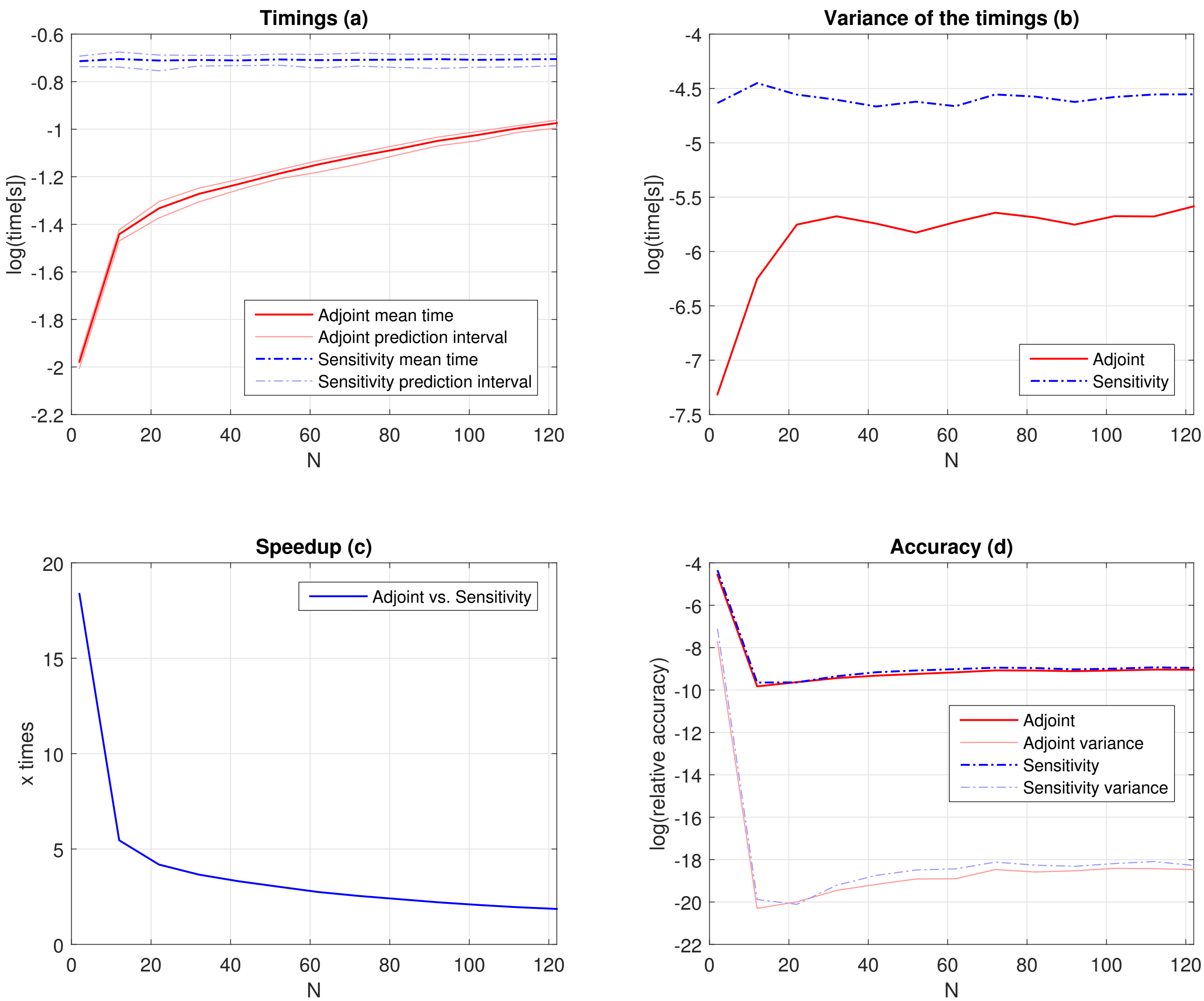

Figure 2: Adjoint gradient for linear ODE model: number of time observations

the second order term (20) can be arbitrarily large with respect to the first order term in (19). This is a well known fact but often overlooked. Let us return to the linear model (24) for a deeper insight. In this case, the first order approximation F of the Hessian of the likelihood H is

$$\begin{aligned} F_{k,k} &= -\sum_{i=1}^{N} e^{\boldsymbol{\phi}_k t_i} e^{\boldsymbol{\phi}_k t_i} t_i^2, \quad k = 1, \ldots, p \\ F_{k,l} &= 0, \quad k \neq l \end{aligned} \tag{28}$$

and the second order term S is

$$\begin{aligned} S_{k,k} &= -\sum_{i}^{N} \left(e^{\boldsymbol{\phi}_k t_i} - \boldsymbol{y}_i \right) e^{\boldsymbol{\phi}_k t_i} t_i^2, \quad k = 1, \ldots, p \\ S_{k,l} &= 0, \quad k \neq l. \end{aligned} \tag{29}$$

We see that the first order approximation F carries absolutely no information about how far the solution is from the data. The Gauss-Newton approximation error can thus be arbitrarily large when $\mathbf{y}$ is not well approximated by the solution $e^{\boldsymbol{\phi} t}$, especially for the values corresponding to small measurement times.

This is for example exploited in the well-known Levenberg-Marquardt method for the least-square minimization (Moré, 1978), which dynamically switches from the gradient descent method (GD) to the Gauss-Newton (GN) method. The GD is used to get sufficiently close to a minimum, so that the GN approximation is reliable.

The overall setup stays identical to the one used for the gradient, i.e. the one corresponding to Figure 1. For convenience, we consider a shorter parameter range $\dim(\boldsymbol{\phi}) \in [2, 52]$, since the finite difference approximation of Hessian, to which we compare the ASM, has quadratic complexity in p. It makes the experiments more time prohibiting in comparison to the gradient. The results are depicted in Figure 3.

FD-adjoint, i.e. approximating hessians using (21), is the most time-efficient approach (Figure 3(a)). The speedup with respect to the finite difference approximation (FD) is linear in $\dim(\boldsymbol{\phi})$ as expected (Figure 3(c)). The FD-adjoint Hessian accuracy is usually sufficient (Figure 3(d)) and moreover it is behaving well as a function of the dimension of the parametric space.

If higher accuracy up to machine precision is desirable, one can compute Hessian using (20), i.e. using SA-approach. Our SA-implementation is however clearly slower than the FD-adjoint despite of a rather optimal coding. FD-adjoint is superior time-wise mainly due to its simplicity.

To conserve space, we do not include any experiments for Hessian with respect to the number of measurement times N. But obviously, for the Hessian computed via FD-adjoint, the results for gradient are directly applicable. We will analyze the dependence on N for the following model in Section 6.2. Here we have focused on $p-$ scaling only, which cannot be tested for the realistic model.

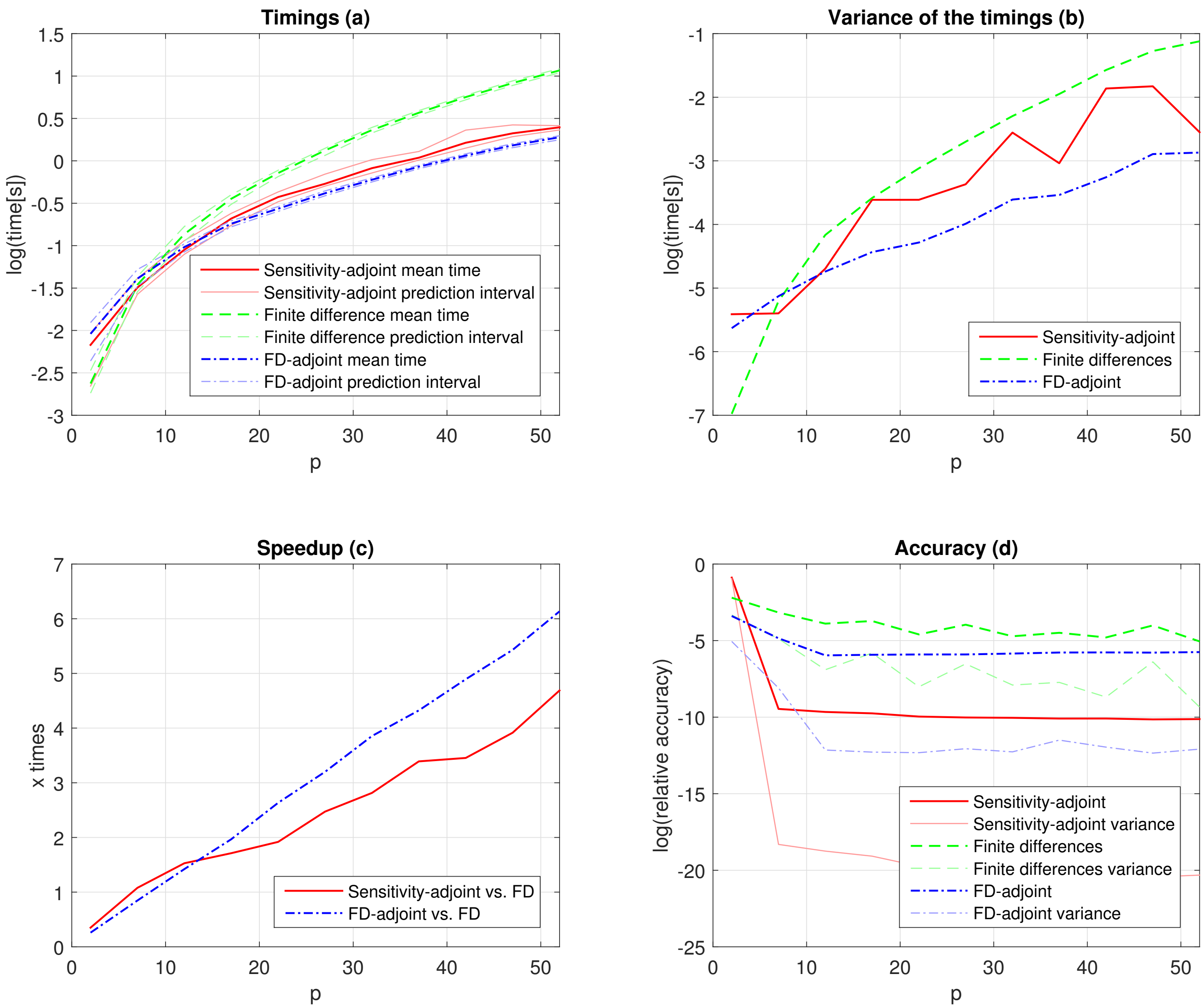

Figure 3: Adjoint Hessian for linear ODE model: parameter space dimension

Table 1: Parameters of the latent dynamic HIV model

$\boldsymbol{\phi}$	λ	γ	μ_{NI}	μ_L	μ_A	μ_V	p	α_L	π	η_{NRTI}	η_{PI}
$\boldsymbol{\phi}_0$	2.61	.0021	.0085	.0092	.289	30	641	1.6e−5	.443	.90	.99

6.2 Latent dynamic HIV model

Now we are going to assess the efficiency and accuracy of ASM for a more realistic model - latent dynamic HIV model from Lavielle et al. (2011):

$$
\begin{aligned}
d_t T_{NI} &= \lambda - (1-\eta_{NRTI})\gamma T_{NI} V_I - \mu_{NI} T_{NI},\\
d_t T_L &= (1-\pi)(1-\eta_{NRTI})\gamma T_{NI} V_I - \alpha_L T_L - \mu_L T_L,\\
d_t T_A &= \pi(1-\eta_{NRTI})\gamma T_{NI} V_I + \alpha_L T_L - \mu_A T_A,\\
d_t V_I &= (1-\eta_{PI}) p T_A - \mu_V V_I,\\
d_t V_{NI} &= \eta_{PI} p T_A - \mu_V V_{NI},
\end{aligned}
\tag{30}
$$

where T_{NI} is the number of not-infected CD4 cells, T_L of latent infected CD4 cells and T_A the number of active infected CD4 cells producing new virons. The number of infectious viruses is V_I and the non-infectious V_{NI}. The 11 parameters $\boldsymbol{\phi}$ represent mostly rates of change. The two of them $\eta_{NRTI}, \eta_{PI} \in [0,1]$ represent the efficacies of two types of antiviral therapies. The available measurements $\boldsymbol{y}_i$ are restricted to the cumulative counts of CD4 cells and the virons, i.e. $V_{ij} = V_I + V_{NI}$ and $T_{ij} = T_{NI} + T_L + T_A$ respectively. For the details see Lavielle et al. (2011). We have $p = 11$, $m = 5$ and $n = 2$.

The setup of the experiments stays rather similar to the previous ones. The parameters are generated randomly around $\boldsymbol{\phi}_0$ which is presented in Table 1 as follows:

$$
\boldsymbol{\phi}_i \sim U[0.95\boldsymbol{\phi}_{0,i}, 1.05\boldsymbol{\phi}_{0,i}], \quad 1 \le i \le p. \tag{31}
$$

The efficacies η_{NRTI} and η_{PI} can be sometimes generated out of the allowed range $[0,1)$. We project them back:

$$
\boldsymbol{\phi}_i = \min(\boldsymbol{\phi}_i, 0.999), \quad i \in \{10, 11\}.
$$

The corresponding synthetic measurements $\boldsymbol{y}$ are perturbed using (27).

For the HIV model p is fixed and we can supply the results only with respect to N. We again observe in Figure 4(a) that the efficiency of (16) is strongly dependent on the number of observations. For up to 5 observations it is more efficient than the sensitivity equation (SE) approach. Thus even for rather small models, the ASM approach for the computation of the gradient of (2) can be advised for certain applications, such as mixed effects modeling in pharmacokinetics and pharmacodynamics, as in (Lavielle et al., 2011). However, for small models with a high number of observation points, the SE approach is clearly more efficient. Accuracy-wise, both approaches are equivalent (Figure 4(d)).

In Figure 5 the corresponding results for the Hessian computation of (2) are presented. Three ways are compared: finite difference (FD) approximation, the ASM approach (SA) using (19) and (20) and the mixed approach (FA) using (21).

First, again as for the diagonal linear model in Section 6.1, the efficiency of the FD approximation is virtually independent of the number of measurements (Figure 5(a)). This is not the case for the SA and FA approaches. However, due to its simplicity, the mixed FA approach is clearly more efficient than SA. It is more efficient than the FD approach up to 5 measurements, which corresponds to the previous results for the computation of the gradient.

The accuracies in Figure 5(d) are compared to the results of SA approach, as no exact solution is available. For the linear model (24), this approach was shown to be accurate up to the machine precision. The mixed FA approach achieves stable accuracies around 10^{-6}, two orders of magnitude better then the full finite difference approximation.

7 Conclusions

We have derived and analyzed the adjoint-state method for computation of the gradient and the Hessian of likelihood functionals for time series data modeled by ordinary differential equations. We interfaced the discrete data and the continuous model on the level of likelihood functional, using the concept of point-wise distributions. The resulting adjoint problem (15) then contains a Dirac delta source corresponding to individual measurement times. The developments are fully supported by the corresponding theoretical results. The implementation of a solver to (15) closely follows the constructive proof of its well-posedness.

Then, we compared the efficiency of the resulting ASM with finite differences and sensitivity equation (SE) approaches, both for the gradient and the Hessian. First, the implementation of SE approach is so efficient, that it renders the finite difference approximation practically obsolete, due to its superior accuracy. Second, the ASM efficiency is dependent on the number of measurement times, which is not the case for SE approach. For models with a high-number of parameters and a small number of measurement times, the ASM is a clear winner. It starts to be competitive even for rather small models like the latent dynamic HIV model from Section 6.2 (11 parameters, 6 measurement times).

In future, we plan a sparse matrix code rewrite, which would allow for solution to bigger ODE systems and also a computationally more efficient implementation. Then, the preconditioning of Newton solver step during the CVODES integration of (15) is an interesting possibility to speed up the ASM.

SUPPLEMENTAL MATERIALS

Appendix A: Appendix containing the proof of the presented results

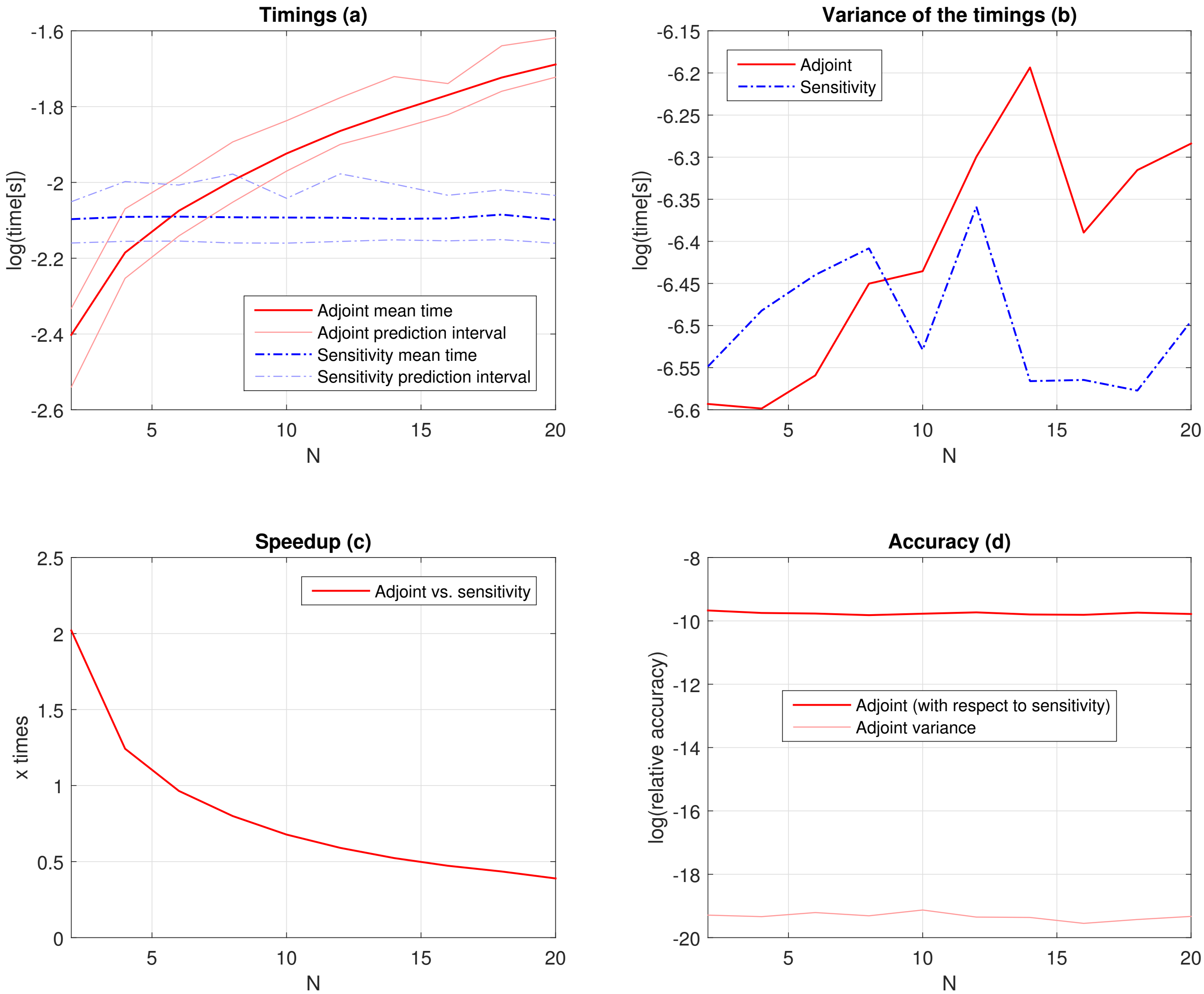

Figure 4: Adjoint gradient for HIV model: number of time observations

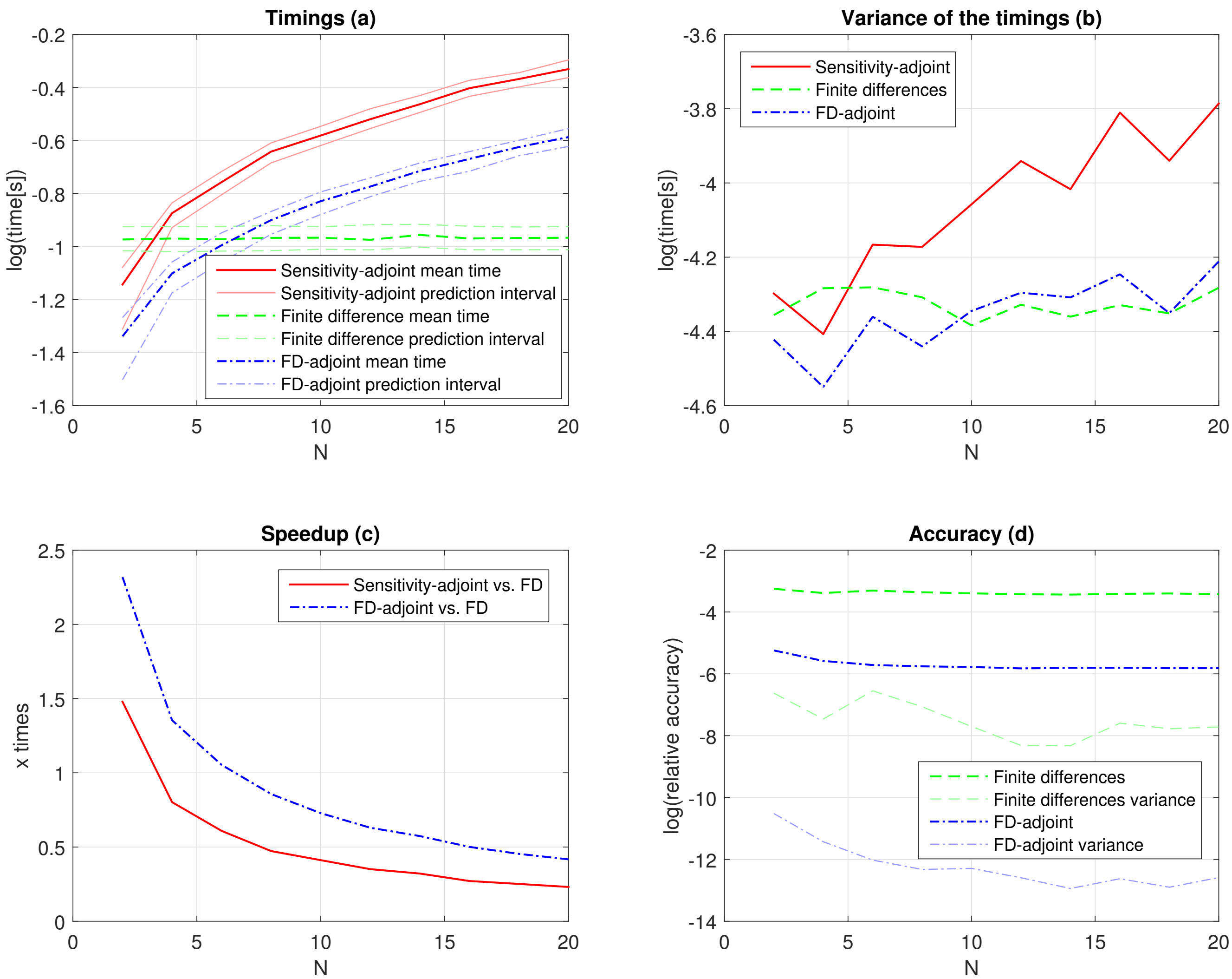

Figure 5: Adjoint Hessian for HIV model: number of time observations

A Proofs

Proof of Theorem 1

First, when compared with Theorem 4.D from Zeidler (1985) we work with $X = \mathbb{R}^m$ and $P = \mathbb{R}^p$. Those are complete normed vector spaces i.e. they are Banach spaces. Second, the initial condition is dependent only on parameter $\boldsymbol{\phi}$, not on any free parameter y as in Theorem 4.D.

Set $J = [-1,1]$. Let us do the following rescaling: $t = sa$, $\boldsymbol{z}(s) := \boldsymbol{u}(as) - \boldsymbol{u}_0(\boldsymbol{\phi})$ for all $s \in J$. Then (1) is equivalent to

$$\boldsymbol{z}'(s) - a\boldsymbol{f}(as, \boldsymbol{z}(s) + \boldsymbol{u}_0(\boldsymbol{\phi}), \boldsymbol{\phi}) = \mathbf{0} \quad \text{for all } s \in J, \boldsymbol{z}(0) = \mathbf{0}.$$

This can be written as an operator equation $F(\boldsymbol{z}, a, \boldsymbol{\phi}) = 0$ with the operator $F : \boldsymbol{Z} \times \boldsymbol{A} \to \boldsymbol{W}$ and spaces $\boldsymbol{Z} = \{\boldsymbol{z} \in \boldsymbol{C}^1(J, \mathbb{R}^m) : \boldsymbol{z}(0) = \mathbf{0}\}$, $\boldsymbol{W} = C(J, \mathbb{R}^m)$. The space $\boldsymbol{A}$ contains all the parameters $(a, \boldsymbol{\phi})$, i.e. $\boldsymbol{A} = \mathbb{R} \times \mathbb{R}^p$.

Set $\boldsymbol{q} = (\mathbf{0}, 0, \boldsymbol{\phi})$. Both F and F_z are continuous at $\boldsymbol{q}$. Obviously, $F(\boldsymbol{q}) = \mathbf{0}$ and $F_z(\boldsymbol{q})\boldsymbol{z} = \boldsymbol{z}'$. The crucial observation is that for every $\boldsymbol{w} \in \boldsymbol{W}$, there exists exactly one $\boldsymbol{z} \in \boldsymbol{Z}$ with $\boldsymbol{z}' = \boldsymbol{w}$, namely $\boldsymbol{z}(s) = \int_0^s \boldsymbol{w}(t)\,\mathrm{d}t$. Hence $F_z(\boldsymbol{q}) : \boldsymbol{Z} \to \boldsymbol{W}$ is bijective. The implicit function theorem yields the conclusions (see e.g. Theorem 4.B in Zeidler (1985)). ■

Proof of Lemma 1

After realizing that l depends on $\boldsymbol{\phi}$ only through $\boldsymbol{u}$, (12) is formally a direct application of the chain rule ($d_{\boldsymbol{u}}$ denotes the derivative of the metric with respect to the model state $\boldsymbol{u}$.)

The r.h.s. of (12) is a well-posed finite expression. First, we have assumed that the metric is sufficiently smooth, thus $d_{\boldsymbol{u}}$ is continuous. Second, $\boldsymbol{s}$ is continuous as well owing to Theorem 1. A distribution can be rescaled by any at least continuous function, as here $\{\delta\}$ by $d_{\boldsymbol{u}}$.

Thus, the first differential on the l.h.s of (12) exists as well. It is moreover continuous, i.e. it is Fréchet. [6] ■

Proof of Theorem 2

First, let us assume that we have already constructed a unique solution $\boldsymbol{v}$ to (15) up to a certain measurement point t_i. The adjoint problem is solved backwards in time. Consequently, we will construct its prolongation on $[t_i, t_{i-1})$.

Let $\boldsymbol{v}_i^+$ be the ODE solution just before integrating the measurement at time t_i, i.e. at time t_i^+. We

[6] An alternative argumentation could employ equivalence (9).

simply stop the integration at t_i^+, add $d_{\boldsymbol{u}}(\boldsymbol{y}_i, \boldsymbol{g}(t_i, \boldsymbol{\phi}))$ to $\boldsymbol{v}_i^+$ and solve

$$\begin{aligned} d_t \boldsymbol{v} &= -J_{\boldsymbol{u}}^*(\boldsymbol{f})\boldsymbol{v}, \quad t \in (t_i, t_{i-1}), \\ \boldsymbol{v}(t_i) &= \boldsymbol{v}_i^+ + d_{\boldsymbol{u}}(\boldsymbol{y}_i, \boldsymbol{g}(t_i, \boldsymbol{\phi})). \end{aligned} \tag{32}$$

This is a simple linear ODE with a continuous coefficient $J_{\boldsymbol{u}}^*(\boldsymbol{f})$, since $f \in \boldsymbol{C}^1$. The classical results yield the global solution on (t_i, t_{i-1}) (see e.g. Theorem 5.1 and Theorem 5.2 from (Coddington and Levinson, 1955).) This concludes the proof of the existence and uniqueness.

Now, we prove (14). Let us without a loss of generality assume that there are no measurements in times 0 and T. It is a well-known result of theory of distributions (in the sense of functional analysis), that the classical integration by part formula

$$\int_0^T d_t \boldsymbol{v}\boldsymbol{w}\, dt = [\boldsymbol{v}\boldsymbol{w}]_0^T - \int_0^T \boldsymbol{v} d_t \boldsymbol{w}\, dt \tag{33}$$

is valid for $\boldsymbol{w} \in \boldsymbol{C}^1$ even if the derivative $d_t \boldsymbol{v}$ exists on $[0, T]$ only in a *weak* sense, i.e. almost everywhere. Actually, (33) is the definition of the weak derivative of $\boldsymbol{v}$ taking only $\boldsymbol{w} \in \boldsymbol{C}_0^1([0,T])$. Consequently, since $\boldsymbol{s} \in \boldsymbol{C}^1([0,T])$, we can safely proceed as follows

$$\begin{aligned} \langle \boldsymbol{s}, \{\delta\} d_{\boldsymbol{u}}(\boldsymbol{y}, \boldsymbol{g}(t, \boldsymbol{\phi}))\rangle &\overset{(15)}{=} \langle \boldsymbol{s}, d_t \boldsymbol{v} + J_{\boldsymbol{u}}^*(\boldsymbol{f})\boldsymbol{v}\rangle \\ &\overset{(8)}{=} -\boldsymbol{v}^*(0) J_{\boldsymbol{\phi}}(\boldsymbol{u}_0)\boldsymbol{h} - (d_t \boldsymbol{s} - J_{\boldsymbol{u}}(\boldsymbol{f})\boldsymbol{s}, \boldsymbol{v}) \\ &\overset{(8)}{=} -\boldsymbol{v}^*(0) J_{\boldsymbol{\phi}}(\boldsymbol{u}_0)\boldsymbol{h} - \left(J_{\boldsymbol{\phi}}(\boldsymbol{f})\boldsymbol{h}, \boldsymbol{v}\right). \blacksquare \end{aligned} \tag{34}$$

Proof of Lemma 2

The existence and uniqueness of $\boldsymbol{\varsigma}$ is a direct results of Theorem 1. Now, (20) is derived as follows:

$$\begin{aligned} \langle \boldsymbol{\varsigma}, \{\delta\} d_{\boldsymbol{u}}(\boldsymbol{y}, \boldsymbol{g}(t, \boldsymbol{\phi}))\rangle &\overset{(15)}{=} \langle \boldsymbol{\varsigma}, d_t \boldsymbol{v} + J_{\boldsymbol{u}}^*(\boldsymbol{f})\boldsymbol{v}\rangle \\ &\overset{(15),(18)}{=} -(\boldsymbol{u}_0)_{\boldsymbol{\phi}\boldsymbol{\phi}} \boldsymbol{h}_1 \boldsymbol{h}_2 \cdot \boldsymbol{v}(0) \\ &\quad - (d_t \boldsymbol{\varsigma}, \boldsymbol{v}) + (J_{\boldsymbol{u}}(\boldsymbol{f})\boldsymbol{\varsigma}, \boldsymbol{v}). \end{aligned} \tag{35}$$

This after substituting for $J_{\boldsymbol{u}}(\boldsymbol{f})\boldsymbol{\varsigma}$ from (18) directly yields (20). Analogically to the proof of Theorem 2, we needed $\boldsymbol{\varsigma} \in \boldsymbol{C}^1([0,T])$ to be able to integrate by parts.$\blacksquare$

References

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006), *Nonlinear programming: theory and algorithms*, John Wiley & Sons, 3rd ed.

Bensoussan, A., Lions, J., and Papanicolaou, G. (1978), *Asymptotic Analysis for Periodic Structures*, Amsterdam: North-Holland Pub. Co.

Bertsekas, D. P. (1999), *Nonlinear programming*, Athena scientific Belmont.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011), *Handbook of Markov Chain Monte Carlo*, CRC Press.

Cesari, L. (1983), *Optimization–theory and applications: problems with ordinary differential equations*, vol. 17, Springer-Verlag.

Cimrák, I. and Melicher, V. (2007), "Sensitivity analysis framework for micromagnetism with application to the optimal shape design of magnetic random access memories," *Inverse Problems*, 23, 563–588.

Coddington, E. A. and Levinson, N. (1955), *Theory of ordinary differential equations*, Tata McGraw-Hill Education.

Delyon, B., Lavielle, M., and Moulines, E. (1999), "Convergence of a stochastic approximation version of the EM algorithm," *The Annals of Statistics*, 27, 94–128.

Draelants, D., Broeckhove, J., Beemster, G. T. S., and Vanroose, W. (2012), "Numerical bifurcation analysis of the pattern formation in a cell based auxin transport model," *Journal of Mathematical Biology*, 67, 1279–1305.

Gillespie, D. T. (1977), "Exact stochastic simulation of coupled chemical reactions," *The journal of physical chemistry*, 81, 2340–2361.

Hindmarsh, A. C. (1980), "LSODE and LSODI, Two New Initial Value Ordinary Differential Equation Solvers," *SIGNUM Newsl.*, 15, 10–11.

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

Lavielle, M., Samson, A., Karina Fermin, A., and Mentré, F. (2011), "Maximum Likelihood Estimation of Long-Term HIV Dynamic Models and Antiviral Response," *Biometrics*, 67, 250–259.

Lewis, J. M., Lakshmivarahan, S., and Dhall, S. (2006), *Dynamic data assimilation: a least squares approach*, vol. 13, Cambridge University Press.

Lions, J. L. (1971), "Optimal control of systems governed by partial differential equations," .

Martin, J., Wilcox, L., Burstedde, C., and Ghattas, O. (2012), "A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion," *SIAM Journal on Scientific Computing*, 34, A1460–A1487.

Melicher, V. and Vrábel', V. (2013), "On a continuation approach in Tikhonov regularization and its application in piecewise-constant parameter identification," *Inverse Problems*, 29, 115008.

Moré, J. J. (1978), "The Levenberg-Marquardt algorithm: Implementation and theory," in *Numerical Analysis*, ed. Watson, G., Springer Berlin Heidelberg, vol. 630 of *Lecture Notes in Mathematics*, pp. 105–116.

Murray, J. D. (2002), *Mathematical Biology I: An Introduction*, vol. 17 of *Interdisciplinary Applied Mathematics*, Springer-Verlag, 3rd ed.

Serban, R. and Hindmarsh, A. C. (2005), "CVODES: the sensitivity-enabled ODE solver in SUNDIALS," in *ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, American Society of Mechanical Engineers, pp. 257–269.

Slodička, M. and Balážová, A. (2010), "Decomposition method for solving multi-species reactive transport problems coupled with first-order kinetics applicable to a chain with identical reaction rates," *Journal of Computational and Applied Mathematics*, 234, 1069–1077, proceedings of the Thirteenth International Congress on Computational and Applied Mathematics (ICCAM-2008), Ghent, Belgium, 7–11 July, 2008.

Tornøe, C. W., Agersø¸, H., Jonsson, E., Madsen, H., and Nielsen, H. A. (2004), "Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in NLME using differential equations," *Computer Methods and Programs in Biomedicine*, 76, 3–40.

Wong, R. (2001), *Asymptotic approximations of integrals*, vol. 34, SIAM.

Zeidler, E. (1985), *Nonlinear Functional Analysis and Its Applications: Fixed point theorems*, Nonlinear Functional Analysis and Its Applications, Springer-Verlag.