SIAM/ASA J. Uncertainty Quantification
Vol. 4, pp. 829–849

Bayesian Analysis of ODEs: Solver Optimal Accuracy and Bayes Factors*

Marcos A. Capistrán†, J. Andrés Christen‡, and Sophie Donnet§

Abstract. In most cases in the Bayesian analysis of ODE inverse problems, a numerical solver needs to be used. Therefore, we cannot work with the exact theoretical posterior distribution but only with an approximate posterior derived from the error in the numerical solver. To compare an approximate posterior distribution with the theoretical one, we propose using Bayes factors (BFs), considering both of them as models for the data at hand. From a theoretical point of view, we prove that the theoretical vs. numerical posterior BF tends to 1, in the same order as the numerical solver used. In practice, we illustrate the fact that for higher order solvers (e.g., Runge–Kutta) the BF is already nearly 1 for step sizes that would take far less computational effort. Considerable CPU time may be saved by using coarser solvers that nevertheless produce practically error-free posteriors. Two examples are presented where nearly 90% CPU time is saved, with all inference results being identical to those obtained using a solver with a much finer time step.

Key words. inverse problems, Bayesian inference, Bayes factors, numerical analysis of ODEs

AMS subject classifications. 65L09, 65L06, 62F15, 62P99

DOI. 10.1137/140976777

1. Introduction.

1.1. Context and issues.

In a comprehensive review of recent publications on the Bayesian analysis of inverse problems it is clear that there is a steady growing interest in the uncertainty quantification approach provided by the Bayesian paradigm. Recent reviews on the Bayesian analysis of inverse problems include Mohammad-Djafari (2006), Kaipio and Fox (2011), Watzenig and Fox (2009), or Woodbury (2011).

In recent years, the use of Bayesian inference has emerged not only in the field of image processing (Cai et al. (2011); Chama et al. (2012); Fall et al. (2011); Kolehmainen et al. (2007); Kozawa, Takenouchi, and Ikeda (2012); Nissinen, Kolehmainen, and Kaipio (2011); Zhu et al. (2011), to mention some recent references), but also in a whole range of emerging application areas (see Calvetti et al. (2006); Cui, Fox, and O'Sullivan (2011); Hazelton (2010); Kaipio and Fox (2011); Keats, Yee, and Lien (2010); Wan and Zabaras (2011), to name but a few).

With this increasing use of the Bayesian paradigm in applied inverse problems, theoret-

*Received by the editors July 11, 2014; accepted for publication (in revised form) May 4, 2016; published electronically July 26, 2016.

http://www.siam.org/journals/juq/4/97677.html

Funding: The research for the first and second authors was supported by Fondo Mixto de Fomento a la Investigación Científica y Tecnológica, CONACYT-Gobierno del Estado de Guajanuato, GTO-2011-C04-168776.

†Centro de Investigación en Matemáticas (CIMAT), A.P. 402, Guanajuato, Gto. 36000, Mexico (marcos@cimat.mx).

‡Corresponding author. Centro de Investigación en Matemáticas (CIMAT), A.P. 402, Guanajuato, Gto. 36000, Mexico (jac@cimat.mx).

§INRA, Unité MIA, Equipe MORSE, Paris 75231, France (sophie.donnet@agroparistech.fr).

ical questions have emerged and some answers have been proposed. We now have access to important theoretical results on the definition of posterior distributions in infinite dimensions and regularity conditions for correct approximate schemes through numerical, finite dimension posteriors (e.g., Schwab and Stuart (2012)) that provide a sound theoretical background to the field.

However, as far as we know, only a handful of publications mentions or uses Bayesian predicting tools, that is, the posterior (predictive) distribution of as yet unobserved variables (Capistrán, Christen, and Velasco-Hernández (2012); Kaipio and Fox (2011); Somersalo, Voutilainen, and Kaipio (2003); Vehtari and Lampinen (2000)), and even fewer consider formally the Bayesian model selection and model comparison tools. In this paper, we aim at highlighting the interest of using Bayes factors (BFs) in the inverse problem context.

Predictive power is always a desirable property of mathematical models and inference, beyond parameter estimation, for model parameters that may or may not have straightforward physical meaning. We believe that comparing forward models as *statistical* models is the way to proceed when predictive power is of primary interest. The Bayesian model comparison and model averaging tools, in particular pairwise model comparison using BFs, are in such cases the main tool to be used in this context (Hoeting et al. (1999)).

In particular, BFs could be used when analyzing the numerical vs. the theoretical versions of the resulting posterior distribution. More precisely, in inverse problems, the forward map is defined as the solution of a system of ODEs (or PDEs, etc.) and represents a complex regressor that is only theoretically defined. The actual usage of the model necessarily involves a numerical solver that includes an approximation error depending on the solver step size h. Therefore, on the one hand, there is an exact statistical model $\mathcal{M}$ relying on the theoretical solution of the forward map and depending on some parameters ϕ. On the other hand, the inference is performed on an approximate model relying on the approximate solution of the ODE, $\mathcal{M}_h$, and depending on the same parameters ϕ. Accordingly we derive a theoretical posterior distribution $P_{\Phi|Y}(\phi|y)$ and an approximate posterior distribution $P^h_{\Phi|Y}(\phi|y)$.

Recently a series of papers (e.g., Schwab and Stuart (2012)) discussed regularity conditions under which $P^h_{\Phi|Y}(\phi|y)$ tends to $P_{\Phi|Y}(\phi|y)$ as the approximation error in the forward map tends to zero, using a suitable metric. A metric comparison (i.e., $||P_{\Phi|Y}(\cdot|y) - P^h_{\Phi|Y}(\cdot|y)||$) is useful in proving the required convergence theorems, but more practical considerations will be needed when evaluating the relative benefits of a numerical approach with a particular solver step size h (for data y).

We claim that $P_{\Theta|Y}(\cdot|y)$ and $P^h_{\Theta|Y}(\cdot|y)$ (or $P^{h_i}_{\Theta|Y}(\cdot|y)$ for any other solver step size h_i) may be compared as *models*: $P_{\Theta|Y}(\cdot|y)$ is the reference model but is computationally expensive and sometimes only theoretically available model, while the approximate $P^{h_i}_{\Theta|Y}(\cdot|y)$, for various solver precisions $h_1, h_2, \ldots$, are the alternative less computationally demanding models. BFs may then be used to establish a sound comparison, to balance predictive power on the one hand vs. solver CPU time on the other, to establish a useful solver precision. We attempt to establish how to approximate the BFs, without having the theoretical reference model and using solely the numerical solver approximation rates.

Xue, Miao, and Wu (2010) address related issues of this problem in the statistical analysis of ODE systems from a frequentist perspective. Under several assumptions—in particular,

global identifiability and sufficient smoothness of the nonlinear least-squares estimator—point estimation is shown to be consistent, and with asymptotic normality, under random designs. They provide additional results showing that the step size h of the solver needs to tend to zero at some rate dependent on the sample size in order to have consistency and normality. However, only in passing do they mention the CPU vs. precision trade-off problem. We contribute to the latter and discuss further related issues in a Bayesian context, for a fixed sample, and in a far weaker setting.

1.2. Notation. Assume that we observe a process $\mathbf{y} = (y_1, \ldots, y_n)$ at the discrete times $t_1, \ldots, t_n \in [0, T[^n$ such that

$$y_i = f(X_\theta(t_i)) + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2), \qquad (\mathcal{M}) \tag{1.1}$$

where X_θ is the solution of the following system of ODEs, namely, the forward map:

$$\frac{dX_\theta}{dt} = F(X_\theta, t, \theta); \quad X_\theta(t_0) = X_0. \tag{1.2}$$

$\theta \in A \subset \mathbb{R}^d$ is a vector of unknown parameters, and $F : \mathbb{R}^p \times [0, T[\times A \mapsto \mathbb{R}^p$ is a known function. In our results we take $\sigma^2 \in S \subset \mathbb{R}^+$ to be unknown.

Assumption 1. *We assume throughout the paper that the function F on the right-hand side of the initial value problem (1.2) follows the regularity conditions of Picard's theorem (see* Süli and Mayers (2003, p. 311)*, for example) for all $\theta \in A$ to ensure the existence of a unique solution in the referred initial value problem. We also assume for the parametric space that A and S are compact sets.*

$f : \mathbb{R}^p \to \mathbb{R}^k$ in (1.1) is the observation function. Many types of observation functions f can be considered, modeling, for instance, the observation of a single component of the p-vector $X_\theta(t)$ or a (linear) combination of the components. In this paper, for the sake of simplicity, we consider a one-dimensional observation problem only, that is, $k = 1$. Generalizations of our results to multivariate observations are possible and will be briefly mentioned in section 6.

In the Bayesian paradigm, any statistical decision from the data $\mathbf{y}$—such as estimation, prediction, or model selection—relies on the likelihood function

$$P_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta, \sigma) = \sigma^{-n}(2\pi)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(X_\theta(t_i)))^2\right\}, \tag{1.3}$$

where $\Phi = (\Theta, \Sigma)$ is a random variable with particular realization $\phi = (\theta, \sigma)$. This expression involves the computation of X_θ, a solution of (1.2). However, except in very simple cases, an explicit expression of the solution is in general not available (although its existence is ensured by the regularity conditions on F). As a consequence, in practice, the system (1.2) is solved using a numerical solver and inference is performed, not on the previous "exact" model but on an approximate model, namely,

$$y_i = f(X^h_\theta(t_i)) + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2), \qquad (\mathcal{M}_h) \tag{1.4}$$

where X_θ^h denotes the approximate solution of (1.2) supplied by the numerical solver (h being a precision parameter of the solver, typically its step size). The new likelihood derived from model $\mathcal{M}_h$ is thus

$$P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma) = \sigma^{-n}(2\pi)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(X^h_\theta(t_i)))^2\right\}.$$

Since in general there is no choice but to use the approximate model in (1.4), there exists a real need to understand and control the error made when working with $P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)$ instead of $P_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)$. As implied above, this is the problem we aim to discuss in this paper.

To that purpose, a first natural choice is to compare the posterior distributions calculated from models ($\mathcal{M}$) and ($\mathcal{M}_h$). Such a study has been proposed by, for example, Donnet and Samson (2007). However, when comparing models in a Bayesian context, the natural tools are BFs. In this work, we recall their importance, propose an efficient way to compute them in this context, and study some theoretical aspects of their calculation when the exact model is not available.

The paper is organized as follows. In section 2 we discuss the choice of a solver and its required properties from a Bayesian inverse problem point of view. In section 3, we develop some practical results on error control on posterior distributions as well as on BFs. Some practical aspects of the Bayesian inference and the calculation of BFs (in this inverse problem context) are discussed in section 4. Our results are illustrated both with a simulation study and with real data in section 5. Finally, a discussion of the paper is presented in section 6.

2. ODE solvers from a Bayesian inverse problem point of view. Bayesian analysis for inverse problems strongly relies on the numerical approximation of the underlying ODE system. One can choose to use a standard (more or less advanced) implemented solver as a black box, in a sense assuming that no approximation is made on the model. However, in our approach, we aim at understanding the influence of this approximation. As a consequence, we are interested in the inherent properties of the numerical solver. When it comes to qualifying a numerical solver, three properties arise, namely, its error (local or global), its stability, and its stiffness.

There is a plethora of numerical methods for solving the initial value problem (1.2). Noteworthy are time-stepping methods based on Taylor approximation of the function F, linear multistep methods, and Runge–Kutta methods. These methods span many orders of local accuracy. Besides the order of accuracy, standard requirements for a numerical method are consistency, convergence, and stability (Iserles (1996); Quarteroni, Sacco, and Saleri (2007)). However, even when these latter conditions hold, a common concern in the numerical solution of the initial value problem (1.2) is error control.

Two types of error may be considered, namely, the local and the global errors (Quarteroni, Sacco, and Saleri (2007)). Let h be the step size of the method. In simple terms define a time grid as $t_{i+1} = t_i + h$ for some fixed $h > 0$, and let $X_{\theta,i}$ be the solver approximation of $X_\theta(t_i)$ ($t_0 = 0$). Local truncation error is the error made in one step of the numerical method, while global error E_i is the difference between the computed solution and the true solution at any

given value of t belonging to the grid (Süli and Mayers (2003, Chap. 12)), that is,

$$E_i = E_h(t_i, \theta) = ||X_\theta(t_i) - X_{\theta,i}||_2.$$

Set $X_\theta^h(t_i) = X_{\theta,i}$. The numerical solver is said to have global error control of order p if

$$\max_{t\in\{t_0,t_1,t_2,\ldots,t_n\}} ||X_\theta(t) - X_\theta^h(t)|| \le C_\theta h^p \tag{2.1}$$

for some constant C_θ independent of h. This global error order control characteristic will be needed in section 3.2 to prove our main result.

The solver to be used to approximate the solution of the ODE at hand needs to be suitable and carefully chosen to achieve acceptable performance over the whole parametric space. Arnold, Calvetti, and Somersalo (2013, 2014) stress this fact precisely in the context of Bayesian inference and favor the use of linear multistep methods, like the Adams–Bashforth or Adams–Moulton methods for nonstiff ODEs or the backward differentiation formulae (BDF) for stiff systems. In our experience we have worked with the LSODA FORTRAN package (which is now available in a series of platforms, including Python-SciPy and R; see Radhakrishnan and Hindmarsh (1993)) that dynamically chooses between precisely the previously mentioned solvers. These solvers also tend to be favored in the recent numerical analysis literature (see Süli and Mayers (2003, Chap. 12), for example).

However, the only crucial requirement to establish our main result, as far as the numerical solver is concerned, is the global error control shown in (2.1). The common Adams-type and BDF methods are designed to be *zero-stable*, which in turn may be proved to possess global error control as in (2.1) (Dahlquist's equivalence theorem; see Süli and Mayers (2003, p. 340)) for $p \ge 4$. To ease our presentation and illustrate our results, we chose two ODE examples which are not particularly stiff, and even a simple Runge–Kutta may be used as solver, for which $p = 4$. In fact, in section 5.1 we also used solvers with $p = 2$ and $p = 1$ (Runge–Kutta order 2 and the Euler method, respectively) to further illustrate our result in Theorem 2.

In the numerical community, "error control" means keeping the local and global errors under a fixed level (beyond the above-mentioned asymptotic results). The local error may be controlled directly, using, for instance, the "Milne device" (Iserles (1996)). However, there are no known general methods to control global error E_n, although some methods exist to estimate it; see, for instance, those relying on adjoint state analysis (Cao and Petzold (2004); Lang and Verwer (2007)). In the results that follow, we do not require an estimation of the global (or local) error, solely the knowledge of the global error order for the solver at hand.

Another important issue in the numerical solution of problem (1.2) is stiffness (Lambert (1991)). Many systems of ODEs modeling real-life phenomena are stiff (Gutenkunst et al. (2007)); however, error control in stiff systems is more complex since variable multistep implicit solvers are commonly required. Therefore, we prefer not to consider stiff systems and take nonstiff systems for the two examples in section 5. We add a comment on stiff systems in relation to the methods presented here in section 6.

The results shown in this paper assume a fixed step method. As previously mentioned, our examples only use the Euler and Runge–Kutta methods (orders $p = 1, 2, 4$, respectively).

3. Theoretical results for Bayesian model comparison in inverse problems. We consider a Bayesian framework, setting a prior distribution $P_\Phi(\theta, \sigma)$ on the unknown parameters. We

denote by $P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$ (respectively, $P_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$) the posterior distribution of the approximate (respectively, exact) model, that is,

$$P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y}) = \frac{P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_\Phi(\theta,\sigma)}{P^h_{\mathbf{Y}}(\mathbf{y})},$$
$$P_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y}) = \frac{P_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_\Phi(\theta,\sigma)}{P_{\mathbf{Y}}(\mathbf{y})},$$

where $P^h_{\mathbf{Y}}(\mathbf{y}) = \int P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_\Phi(\theta,\sigma)d\theta d\sigma$ and $P_{\mathbf{Y}}(\mathbf{y}) = \int P_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_\Phi(\theta,\sigma)d\theta d\sigma$ are the normalization constants of the two models, also called the *marginal likelihoods* of data $\mathbf{y}$.

In the next section we discuss the convergence rate of the approximate to the theoretical posterior. In section 3.2 we compare both models using BFs.

3.1. Error control of the approximate posterior distribution. A basic related result may be found in Donnet and Samson (2007) comparing the posterior distributions of the exact and approximate models, respectively—namely, $P_{\Phi|\mathbf{Y}}$ and $P^h_{\Phi|\mathbf{Y}}$—through the total variation distance.

Theorem 1. *Assume that $\phi = (\theta,\sigma)$ remains in a compact set $A \times S$ and that the numerical scheme of step size h is such that $\{t_1,\dots,t_n\} \subset h\mathbb{N}$ and*

$$\max_{t\in\{t_1,\dots,t_n\}} \|X_\theta(t) - X^h_\theta(t)\|_{\mathbb{R}^p} \le C_\theta h^p. \tag{3.1}$$

Also assume that the observation function f is differentiable with a bounded derivative. Then there exists a constant $C_{\mathbf{y}}$ such that for every (θ,σ) and h small enough

$$|P_{\Phi|\mathbf{Y}}(\theta,\sigma;\mathbf{y}) - P^h_{\Phi|\mathbf{Y}}(\theta,\sigma;\mathbf{y})| \le C_{\mathbf{y}} P_\Phi(\theta,\sigma)h^p. \tag{3.2}$$

As a consequence,

$$||P_{\Phi|\mathbf{Y}} - P^h_{\Phi|\mathbf{Y}}||_{TV} \le C_{\mathbf{y}} h^p, \tag{3.3}$$

where $||\cdot||_{TV}$ is the total variation metric. Moreover, there exists another constant $C'_{\mathbf{y}}$ such that

$$\|(\hat\theta^{L^2},\hat\sigma^{L^2}) - (\hat\theta^{h,L^2},\hat\sigma^{h,L^2})\| \le h^p C'_{\mathbf{y}}, \tag{3.4}$$

where $\hat\theta^{L^2},\hat\sigma^{L^2}$ and $\hat\theta^{h,L^2},\hat\sigma^{h,L^2}$ are the posterior expectations of Θ and Σ in the exact and approximate models, respectively.

Proof. The results of Donnet and Samson (2007) were developed for nonlinear mixed effects models in a maximum likelihood context but may be adapted to models (1.1) and (1.4). Inequality (3.2) is derived from Theorem 4 of Donnet and Samson (2007). The error control on the total variation distance is derived directly. Inequality (3.4) is obtained as

follows:

$$\begin{aligned}
&\left\|(\hat{\theta}^{L^2},\hat{\sigma}^{L^2})-(\hat{\theta}^{h,L^2},\hat{\sigma}^{h,L^2})\right\| \\
=&\left\|\int(\theta,\sigma)P_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})d\theta d\sigma-\int(\theta,\sigma)P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})d\theta d\sigma\right\| \\
=&\left\|\int(\theta,\sigma)\left[P_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})-P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})\right]d\theta d\sigma\right\| \\
\leq&\int\|(\theta,\sigma)\|C_{\mathbf{y}}P_{\Phi}(\theta,\sigma)h^p d\theta d\sigma \\
\leq& h^p C_{\mathbf{y}}\int\|(\theta,\sigma)\|P_{\Phi}(\theta,\sigma)d\theta d\sigma=h^p C'_{\mathbf{y}}.
\end{aligned}$$

■

Therefore the numerical posterior approximates the theoretical posterior at the same rate of the solver order. Cotter, Dashti, and Stuart (2010) also presents error control results of the same type.

Even if these results provide interesting theoretical insights, they rely on unknown constants and therefore cannot be used as such in practice. In the next section, we adopt a BF point of view and highlight that such an approach leads to results of more practical interest.

3.2. Comparing the exact and approximate models through BFs. In the Bayesian paradigm, model selection is performed using the BFs defined as follows (see Kass and Raftery (1995) for details). Let $\mathbf{y}$ be the observed data, and let $\mathcal{M}$ and $\mathcal{M}_h$ be our two models in competition (defined in (1.1) and (1.4)).

Consider a prior distribution on the set of the models $\{\mathcal{M},\mathcal{M}_h\}$; the decision between the competing models $\mathcal{M}$ and $\mathcal{M}_h$ is based on the ratio of the posterior probability for each model, namely, the BF

$$B_{\mathcal{M},\mathcal{M}_h}=\frac{P(\mathcal{M}|\mathbf{y})}{P(\mathcal{M}_h|\mathbf{y})}=\frac{P_{\mathbf{Y}}(\mathbf{y})}{P^h_{\mathbf{Y}}(\mathbf{y})}\frac{P(\mathcal{M})}{P(\mathcal{M}_h)},$$

where $P^h_{\mathbf{Y}}(\mathbf{y})$ and $P_{\mathbf{Y}}(\mathbf{y})$ are the integrated likelihoods, or marginal distributions, of $\mathbf{y}$ from model $\mathcal{M}_h$ and $\mathcal{M}$, respectively, defined by

$$\begin{aligned}
P^h_{\mathbf{Y}}(\mathbf{y})&=\int P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi)P_{\Phi}(\phi)d\phi \quad \text{and} \\
P_{\mathbf{Y}}(\mathbf{y})&=\int P_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi)P_{\Phi}(\phi)d\phi.
\end{aligned}$$

As above, $P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi)$ and $P_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi)$ are the approximate and theoretical likelihoods, respectively, and $P_{\Phi}(\phi)$ is the (common) prior distribution for the parameters $\Phi=(\Theta,\Sigma)$.

$\int P_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_{\Phi}(\theta,\sigma)d\theta d\sigma$ has in general no closed form, and several methods have been proposed to estimate it (see section 4). However, all the classical methods (such as Monte Carlo or Markov chain Monte Carlo) require the computation of the likelihood function, which is not available in the theoretical model $\mathcal{M}$. To tackle this point, we present our main result in the following theorem.

Theorem 2. *Assume that the numerical solver is such that the global error may be written as*

$$\max_{t\in\{t_0,t_1,t_2,\ldots,t_n\}} ||X_\theta(t) - X_\theta^h(t)|| \leq C_\theta h^p, \tag{3.5}$$

where h is the step size of the method. In addition, assume that the observation function f is differentiable on $\{X_\theta(t), \theta \in \Theta, t \in [0,T]\}$. Then $P_{\mathbf{Y}}(\mathbf{y}) = P_{\mathbf{Y}}^h(\mathbf{y}) + O(h^p)$. That is, there exists a constant $B(\mathbf{y}) \in \mathbb{R}$ (which does not depend on h) such that

$$\frac{P_{\mathbf{Y}}(\mathbf{y})}{P_{\mathbf{Y}}^h(\mathbf{y})} \simeq 1 + B(\mathbf{y})h^p.$$

Proof. Using the asymptotic behavior of the global error truncation in (3.5) and assuming that f is differentiable, we can write

$$D_h(t,\theta) = f(X_\theta^h(t)) - f(X_\theta(t)) = \nabla f\left(X_\theta(t)\right)\left(X_\theta^h(t) - X_\theta(t)\right) + O(h^{2p}) = O(h^p). \tag{3.6}$$

This approximation allows us to obtain a development of the marginal likelihood. Let $R_h(\phi) = \frac{P_{\mathbf{Y}|\Phi}^h(\mathbf{y}|\theta,\sigma)}{P_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)}$. Then

$$\begin{aligned} P_{\mathbf{Y}}^h(\mathbf{y}) &= \int P_{\mathbf{Y}|\Phi}^h(\mathbf{y}|\phi)P_\Phi(\phi)d\phi = \int P_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi)R_h(\phi)P_\Phi(\phi)d\phi \\ &= P_{\mathbf{Y}}(\mathbf{y}) + \int P_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi)(R_h(\phi)-1)P_\Phi(\phi)d\phi. \end{aligned}$$

We see that

$$\begin{aligned} R_h(\phi) - 1 = \exp\Bigg\{ &-\frac{1}{2\sigma^2}\sum_{i=1}^n \left[f(X_\theta^h(t_i)) - f(X_\theta(t_i))\right]^2 \\ &+ 2\left[y_i - f(X_\theta(t_i))\right]\left[f(X_\theta(t_i)) - f(X_\theta^h(t_i))\right]\Bigg\} - 1 \\ = &-\frac{1}{2\sigma^2}\sum_{i=1}^n D_h(t_i,\theta)^2 + 2(y_i - f(X_\theta(t_i)))D_h(t_i,\theta) + O(D_h(t_i,\theta)^2), \end{aligned} \tag{3.7}$$

since $e^x - 1 = x + O(x^2)$ for x small enough. Using the expression in (3.6) for D_h, we must have that $R_h(\phi) - 1 = O(h^p)$. From this (and since, from Assumption 1, the parameter space is compact) we get $P_{\mathbf{Y}}^h(\mathbf{y}) = P_{\mathbf{Y}}(\mathbf{y}) + O(h^p)$. Indeed, this also implies $P_{\mathbf{Y}}(\mathbf{y}) = P_{\mathbf{Y}}^h(\mathbf{y}) + O(h^p)$, and therefore for h small enough,

$$\frac{P_{\mathbf{Y}}(\mathbf{y})}{P_{\mathbf{Y}}^h(\mathbf{y})} \simeq 1 + B(\mathbf{y})h^p.$$

■

Corollary 1. *If* $\hat{g} = \int g(\phi) P_{\Phi|\mathbf{Y}}(\phi|\mathbf{y}) d\phi$ *and* $\hat{g}^h = \int g(\phi) P^h_{\Phi|\mathbf{Y}}(\phi|\mathbf{y}) d\phi$ *exist, then*

$$|\hat{g}^h - \hat{g}| = \frac{P_{\mathbf{Y}}(\mathbf{y})}{P^h_{\mathbf{Y}}(\mathbf{y})} B_g(\mathbf{y}) h^p = O(h^p)$$

for some constant $B_g(\mathbf{y})$.

Proof. Note that $|\hat{g}^h - \hat{g}| = \left| \int g(\phi) R_h(\phi) \frac{P_{\mathbf{Y}}(\mathbf{y})}{P^h_{\mathbf{Y}}(\mathbf{y})} P_{\Phi|\mathbf{Y}}(\phi|\mathbf{y}) d\phi - \hat{g} \right|$ and therefore

$$|\hat{g}^h - \hat{g}| = \frac{P_{\mathbf{Y}}(\mathbf{y})}{P^h_{\mathbf{Y}}(\mathbf{y})} \left| \int g(\phi)(R_h(\phi) - 1) P_{\Phi|\mathbf{Y}}(\phi|\mathbf{y}) d\phi - \left(\frac{P^h_{\mathbf{Y}}(\mathbf{y})}{P_{\mathbf{Y}}(\mathbf{y})} - 1 \right) \hat{g} \right|.$$

Combining (3.7) and the above theorem, one reaches the result. ∎

We comment on the above result:

• From (3.7), we note that the error in the regression term $D_h(t, \theta)$ is not important per se except in respect to the observation noise standard error σ. This obvious remark has consequences. It means that when working on a statistical model involving the numerical approximation of a differential system, there is no need for choosing a step size as small as possible; it is enough simply to adapt it such that the global error is small with respect to σ. This can allow computational time savings, as illustrated by the numerical examples presented in section 5.

• Note that $B(\mathbf{y})$ depends only on the numerical method and on the data, but not on the step size h.

• The marginal likelihood $P_{\mathbf{Y}}(\mathbf{y})$ of the unavailable theoretical model now may be estimated. Indeed, an obvious method is to compute $P^{h_k}_{\mathbf{Y}}(\mathbf{y})$ for various step sizes $\{h_k, k = 1, \ldots, K\}$ and fit the simple linear regression $P^{h_k}_{\mathbf{Y}}(\mathbf{y}) = a + b h_k^p$; a then provides an estimation of $P_{\mathbf{Y}}(\mathbf{y})$. This means that by using a multiresolution computation of $P^{h_k}_{\mathbf{Y}}(\mathbf{y})$ on various approximate models, we are able to estimate the marginal likelihood of the true model. We use this procedure to illustrate our results in section 5.

In the next section we discuss the calculation of integrated likelihoods in this context. In section 5 we develop two examples where we illustrate our results.

4. Computation of BFs in an inverse problems context.

There are some excellent reviews concerning the Bayesian analysis of inverse problems (Fox, Palm, and Nicholls (1999); Kaipio and Somersalo (2005)), and for a more detailed description, these or other sources should be consulted. Here we only present the particular aspects of the field relevant to the computation of the BFs.

4.1. Sampling the posterior distribution $P^h_{\Phi|\mathbf{Y}}(\theta, \sigma|\mathbf{y})$.

Basically only for conjugate models the posterior distribution $P^h_{\Phi|\mathbf{Y}}(\theta, \sigma|\mathbf{y})$ has a known form and the normalizing constant $P^h_{\mathbf{Y}}(\mathbf{y})$ (the integrated likelihood) may be easily calculated. Otherwise, $P^h_{\mathbf{Y}}(\mathbf{y})$ needs to be calculated numerically.

If the dimension of Φ is 1 or 2, we could rely on numerical integration to obtain the normalizing constant $P^h_{\mathbf{Y}}(\mathbf{y})$. In larger dimensions, the standard solution is to resort to Monte Carlo methods to sample from the posterior distribution at hand. Let $(\theta^{(l)}, \sigma^{(l)})_{l=1,\ldots,M}$ be a sample

from the posterior distribution $P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$; the posterior mean estimator, for instance, is approximated as $(\hat{\theta}^{L^2},\hat{\sigma}^{L^2}) = \left(\frac{1}{M}\sum_{l=1}^M \theta^{(l)}, \frac{1}{M}\sum_{l=1}^M \sigma^{(l)}\right)$.

Simulation from the posterior distribution is not a direct task, and Markov chain Monte Carlo (MCMC) algorithms (Robert and Casella (2004)) are standard tools to sample from the posterior distribution $P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$. MCMC is especially suited for sampling from complex multidimensional distributions and is ubiquitous in modern Bayesian analyses (Robert and Casella (2004)). The principle of MCMC algorithms is to generate a Markov chain whose invariant distribution is the distribution of interest $P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$. Many versions have been proposed in the literature. Among them, the Gibbs algorithm and the Metropolis–Hastings (MH) algorithms are the most used. While the Gibbs sampler is a very popular MCMC algorithm, it makes sense to use it only in some particular cases (when the full conditionals have a known form). In our inverse problem setting, this is not the case, and the general MH MCMC algorithm needs to be used instead (Robert and Casella (2004)). In this algorithm, assume that the chain has reached the value $\phi^{(\ell)}$; a new parameter value ϕ' is proposed with a proposal distribution $q(\phi'|\phi^{(\ell)})$ and is accepted with probability

$$\rho(\phi',\phi^{(\ell)}) = \min\left\{1, \frac{P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi')P_\Phi(\phi')}{P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi^{(\ell)})P_\Phi(\phi^{(\ell)})}\frac{q(\phi^{(\ell)}|\phi')}{q(\phi'|\phi^{(\ell)})}\right\}. \tag{4.1}$$

MCMC methods tend to be very intensive, commonly with many rejected steps. As a consequence, in the particular case of inverse problems, these methods are even more difficult to use since, *at each iteration of the MH-MCMC*, the forward map needs to be evaluated (in order to compute the likelihood in the MH acceptance probability). It is then crucial to minimize the number of iterations in the MCMC to be used.

Optimizing MCMC algorithms (that is to say, minimizing the number of iterations) has been a very active research topic in the last decade. There are adaptive algorithms (Atchadé and Rosenthal (2005); Haario, Saksman, and Tamminen (1998)) that require additional regularity conditions on the adaptive scheme, model, and prior that might limit their applicability. Christen and Fox (2010) also propose the t-walk, which self-adjusts, keeping two points in the parameter space, and that commonly samples with reasonable efficiency. However, robust, multipurpose, automatic, and optimal methods are still far away on the MCMC horizon (to make an optimistic metaphor).

An initial straightforward but very useful computational economy is to save the $U_\ell = -\log P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\phi^{(\ell)}) - \log P_\Phi(\phi^{(\ell)})$ at each evaluation of the unnormalized posterior within the MCMC scheme. Indeed, this quantity will be used for any new simulated value ϕ' until the chain accepts a new point. We will see in the following subsection that these quantities can also be recycled for model comparison purposes and the calculation of BFs.

Recently Arnold, Calvetti, and Somersalo (2013, 2014) advocated the use of sequential Monte Carlo approaches to this problem, with the potential advantage that the ODE is sequentially solved ("stirred") over a progressive time interval, over a population of parameter points. These methods may indeed present an alternative to standard MCMC methods.

4.2. Computation of the Bayes factor. We now consider the calculation of the marginal likelihood $P^h_{\mathbf{Y}}(\mathbf{y})$ involved in the BF. In our inverse problem context—where each iteration of

the MCMC requires the computationally intensive approximation of an ODE—we would like to avoid increasing the computational burden by using a specific MCMC and would prefer recycling the output of the MCMC algorithm into a Monte Carlo strategy. An answer can be found in the Gelfand and Dey estimator (Gelfand and Dey (1994)).

Assume that $P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$ is sampled using an intensive Monte Carlo procedure, typically an MH MCMC. Assume that the prior distribution P_Φ is absolutely continuous with respect to the Lebesgue measure, and thus $P_\Phi(\theta,\sigma)$ and $P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$ are densities in the usual sense. The Gelfand and Dey estimator relies on the expression

$$\left[P^h_{\mathbf{Y}}(\mathbf{y})\right]^{-1} = \int \frac{\alpha(\theta,\sigma)}{P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_\Phi(\theta,\sigma)} P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})d\theta d\sigma,$$

where α is any density ($\int \alpha(\theta,\sigma)d\theta d\sigma = 1$) with support containing the support of the posterior.

Now, let $\theta^{(1)},\sigma^{(1)},\theta^{(2)},\sigma^{(2)},\ldots,\theta^{(L)},\sigma^{(L)}$ be an MCMC sample of the posterior $P^h_{\Phi|\mathbf{Y}}(\theta,\sigma|\mathbf{y})$. Considering the above expression, the desired marginal may be approximated by

$$\widehat{P}^h_{\mathbf{Y}}(\mathbf{y}) = \left[\frac{1}{L}\sum_{l=1}^{L} \frac{\alpha(\theta^{(l)},\sigma^{(l)})}{P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta^{(l)},\sigma^{(l)})P_\Phi(\theta^{(l)},\sigma^{(l)})}\right]^{-1}. \tag{4.2}$$

The choice of α conditions the quality of the estimator (its variance). If $\alpha(\theta,\sigma) = \pi(\theta,\sigma)$, the estimator $\widehat{P}^h_{\mathbf{Y}}(\mathbf{y})$ is the harmonic mean, which is known to have a dramatic unstable behavior (infinite variance) in some cases. Best strategies are those that use a weighting density α that stabilizes this estimator, for instance, somehow using an α that resembles $P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_\Phi(\theta,\sigma)$. A simple calculation leads to the fact that using a *thinner* tailed α (as opposed to the result in importance sampling) is better suited to obtaining a finite variance for the estimator above (beyond inverse problems, Valpine (2008) discusses further strategies to reduce this variance).

Moreover, in an inverse problem context, it is critical to avoid recalculating the likelihood $P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)$ since it involves numerically solving the ODE system in (1.2). As a consequence we propose proceeding as follows:

- At each iteration of a typical MH MCMC, the computation in (4.1) requires evaluating $P^h_{\Phi|\mathbf{Y}}(\mathbf{y}|\theta^{(l)},\sigma^{(l)})P_\Phi(\theta^{(l)},\sigma^{(l)})$. After the burn-in period, we save these values, letting $U_l = -\log P^h_{\Phi|\mathbf{Y}}(\mathbf{y}|\theta^{(l)},\sigma^{(l)}) - \log P_\Phi(\theta^{(l)},\sigma^{(l)})$.
- A small subsample of $\theta^{(l)},\sigma^{(l)}$, typically of size less than 1,000, is then used to create a kernel density estimate (KDE), which we will use as our weighting density α. This KDE is (typically) a mixture of Gaussians, with support in the whole space, and will roughly resemble the posterior $P^h_{\Phi|\mathbf{Y}}$.
- Let $A_l = -\log \alpha(\theta^{(l)},\sigma^{(l)})$. Then our estimate becomes

$$P^h_{\mathbf{Y}}(\mathbf{y}) \approx \left[\frac{1}{L}\sum_{l=1}^{L}\exp(U_l - A_l)\right]^{-1}.$$

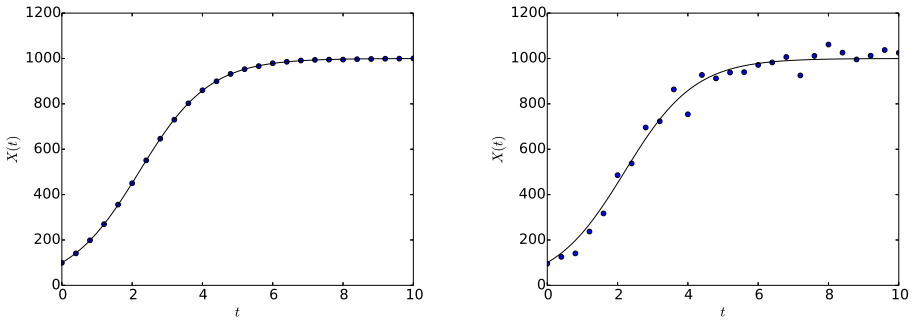


Figure 1. *Synthetic data for the logistic growth with* $\lambda = 1$, $K = 1000$, *and* $\sigma = 1$ *(left) or* $\sigma = 30$ *(right).*

This procedure is fast and basically is a by-product of the MCMC sample, with little CPU burden added. There are robust and fast KDE routines available in popular programming languages like R and Python-Scipy to efficiently produce the weighting function α.

Note that for the Gelfand and Dey estimator, $P^h_{\mathbf{Y}|\Phi}(\mathbf{y}|\theta,\sigma)P_\Phi(\theta,\sigma)$ needs to be known exactly, and coded accordingly. This might be particularly difficult in some situations as, for example, when the prior is not normalized and only implicitly defined. In such situations other estimators may be used (Valpine (2008)), which rely only on ratios of the unnormalized posteriors, although perhaps at a higher computational burden. This is not the case for the examples presented in the next section, and consequently we do not discuss this issue further.

5. Numerical examples.

Here we present three examples to illustrate our results. All examples were programmed in Python-SciPy (python.org, scipy.org) and run on a MacBook Pro laptop with Intel core i7, 2.3 GHz processor.

5.1. Logistic growth models.

We base our first numerical study on the logistic growth model, which is a common model of population growth in ecology. Recently it has also been used to model tumors growth in medicine, among many other applications (Foryś and Marciniak-Czochra (2003)). Let $X(t)$ be the size of the tumor at time t. The dynamics are governed by the differential equation

$$\frac{dX}{dt} = \lambda X(t)(K - X(t)), \quad X(0) = X_0, \tag{5.1}$$

with λK being the growth rate and K the carrying capacity, e.g., $\lim_{t\to\infty} X(t) = K$. Equation (5.1) has an explicit solution equal to

$$X(t) = \frac{KX_0e^{\lambda Kt}}{K + X_0(e^{\lambda Kt} - 1)}.$$

We simulate two synthetic data sets with the error model $y_i = X(t_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and the parameters $X(0) = 100$, $\lambda = 1$, $K = 1000$, $\sigma = 1$ or 30. The data sets are plotted on Figure 1 for the two chosen values of σ. We consider 26 observations at times t_i regularly spaced between 0 and 10.

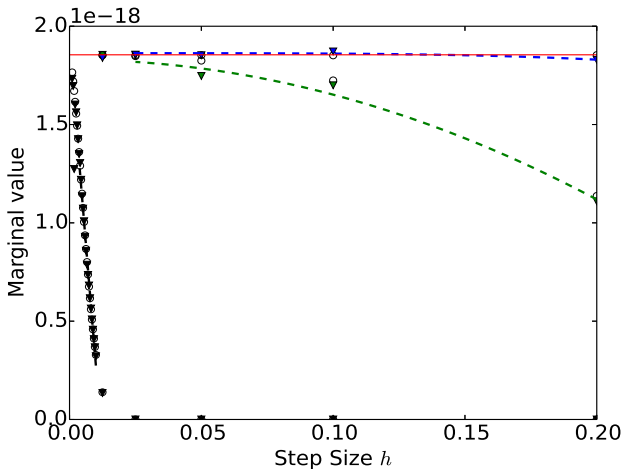


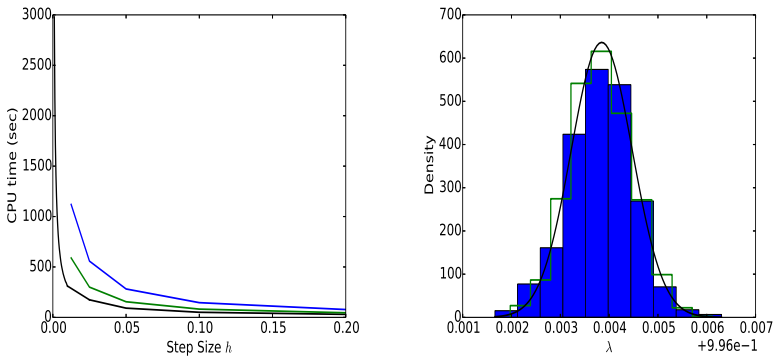
Figure 2. *Study on synthetic data for the logistic growth with $\sigma = 1$. The true marginal $P_{\mathbf{Y}}(\mathbf{y})$ calculated using numerical integration on the analytic solution is represented by the thin horizontal line (red). The marginal $P^h_{\mathbf{Y}}(\mathbf{y})$ for various step sizes is shown as computed by numerical integration (circles) or estimated using the MCMC sample (triangles). Dashed lines indicate the regression for estimated values for $\hat{P}^h_{\mathbf{Y}}(\mathbf{y}) = a + bh^p$ for the orders $p = 1$ (black line on the left), 2 (green), and 4 (blue).*

For this example, K is taken as known, and inference is concentrated on the single parameter λ; since it is a positive parameter we consider a Gamma distribution for the prior on λ.

To highlight our result presented in section 3, we consider the following strategy. For $\sigma = 1$, we first compute what we call the "true" marginal likelihood $P_{\mathbf{Y}}(\mathbf{y})$ (horizontal thin line in Figure 2), using the explicit solution of (5.1) and numerical integration (using the unidimensional quadrature *quad* function from the *scipy.integrate* module). In a second step, we approximate the solution of (5.1) by the Euler scheme (order 1), and Runge–Kutta solvers of orders 2 and 4 (respectively, RK2 and RK4), for various step sizes h_k. The marginal likelihood $P^{h_k(1)}_{\mathbf{Y}}(\mathbf{y})$ is computed using numerical integration and also using the Monte Carlo strategy presented in section 4. These results are plotted with circles for the numerical integration and triangles for the Monte Carlo computation. In the end, for each order p, the estimated values $\hat{P}^{h_k(p)}_{\mathbf{Y}}(\mathbf{y})$ are used to compute the regression functions $\hat{P}^{h_k(p)}_{\mathbf{Y}}(\mathbf{y}) = a + bh^p$ (dashed lines in Figure 2). These results are presented in Figure 2.

The same is done for $\sigma = 30$, but only the RK4 solver is considered; these results are presented in Figure 4. The samples from the posterior distributions are obtained using the t-walk MCMC algorithm (Christen and Fox (2010)). Next we discuss some aspects of this numerical experiment.

• We would like to highlight that the circles and their corresponding triangles in Figures 2 and 4 are quite similar, meaning that the Monte Carlo strategy (derived as a by-product of the MCMC implementation) is an efficient solution to estimating the marginal likelihood in this context. This approximation does not require additional evaluations of the forward map, that is, additional ODE solver runs, and thus has a minimal computational cost.



(a) (b)

Figure 3. (a) *CPU time for various step values h_k and Runge–Kutta solver for orders $p = 1$ (black), $p = 2$ (green), and $p = 4$ (blue), relative to* 10,000 *iterations of the MCMC increasing with order, and exponentially increasing with the step size.* (b) *Posterior distribution of λ for RK4 solver, $p = 4$, for step sizes $h = 0.01$ (blue histogram) and $h = 0.05$ (green histogram) and exact posterior (black density).* 10,000 *iterations of the MCMC took* 17 *minutes for $h = 0.01$ and* 2 *minutes for $h = 0.05$; a* 90% *reduction in CPU time was obtained with no noticeable difference in the resulting posterior distribution.*

• As predicted, the Euler scheme has a linear approximation regime to the correct marginal. To have any substantial savings in CPU time without compromising posterior inference precision, we would need to have a very small step size; i.e., there is no "flat part" in order to take considerable larger step sizes. On the contrary, the RK2 solver, and especially the classical RK4 solver, indeed have a clear flat section where a nearly perfect estimation has been reached. This allows for choosing a much larger step size, meaning a far coarser ODE numerical solver that still makes basically no difference in the resulting inference. This result may be seen in the resulting posterior distributions in Figures 3(b) and 4(c) comparing a very fine solver step size vs. a much larger one, reaching basically the same posterior while utilizing nearly 90% less CPU time.

• For the RK4 solver, we perform a linear regression using the estimated $\hat{P}_{\mathbf{Y}}^{h_k}(\mathbf{y})$ with $h_k = 0.2, 0.1, 0.05, 0.025$ for $\sigma = 1$ and $h_k = 0.8, 0.6, 0.4, 0.2, 0.1, 0.05$ for $\sigma = 30$. Using the formula given in Theorem 2, we deduce an estimation (projection) of the exact marginal likelihood $\hat{P}_{\mathbf{Y}}(\mathbf{y})$, which has to be compared to the true value $P_{\mathbf{Y}}(\mathbf{y})$ (obtained using the exact solution of the ODE and a numerical integration). The results are given in Table 1.

• We believe that an important message is that, for the RK4 solver, as soon as h is lower than some threshold (we took 0.05 for $\sigma = 1$ and 0.1 for $\sigma = 30$), the BF $P_{\mathbf{Y}}^{h_k(2)}(\mathbf{y})/P_{\mathbf{Y}}(\mathbf{y})$ is greater than 0.99, making the models indistinguishable on the Jeffreys scale (Jeffreys (1961, p. 432)) and leading to nearly identical posterior distributions (see Figures 3(b) and 4(c)) for λ. However, the computational time required to estimate the parameters using the smallest h explodes (see Figures 3(a) and 4(b)) from 2 minutes for $h = 0.05$ to 17 minutes for $h = 0.01$, for $\sigma = 1$, and from 2.5 minutes for $h = 0.1$ to 36 minutes for $h = 0.000625$, for $\sigma = 30$. Moreover, for $\sigma = 1$ and for step size $h = 0.02$, the BF $P_{\mathbf{Y}}^{h}(\mathbf{y})/P_{\mathbf{Y}}(\mathbf{y})$ is only 0.61 for the RK2 but already 0.98 for the RK4 solver; see Figure 2.

• Note that the ranges of considered values for h_k are different for $\sigma = 1$ and $\sigma = 30$ (the

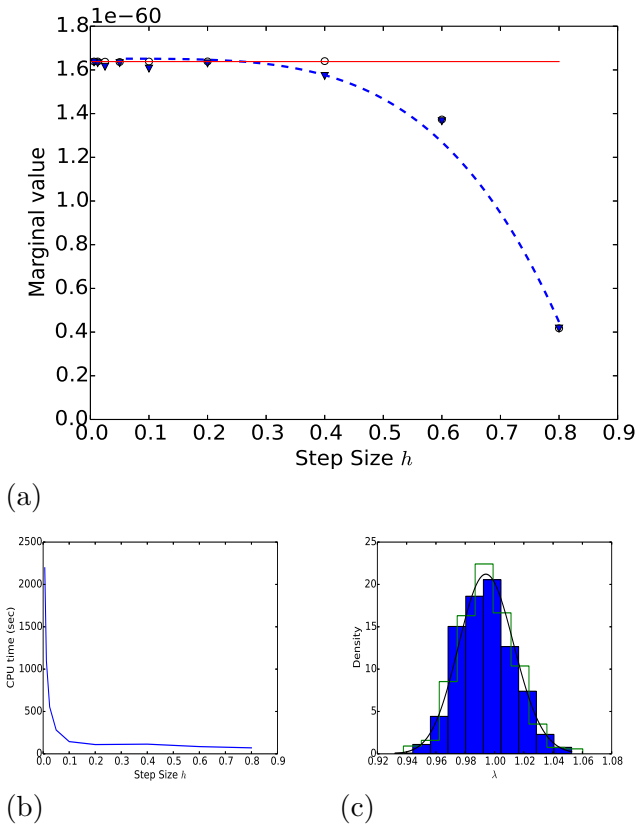


Figure 4. *Study on synthetic data for the logistic growth with* $\sigma = 30$. (a) *Marginal* $P^h_{\mathbf{Y}}(\mathbf{y})$ *for various step sizes, both exact (circles, using numerical integration) and estimated using the MCMC sample (triangles). We use a Runge–Kutta solver of order* 4 *(classical RK4, blue) only. The horizontal line (red) is the true marginal* $P_{\mathbf{Y}}(\mathbf{y})$ *calculated using numerical integration on the analytic solution. Dashed lines indicate the regression for estimated values for* $\hat{P}^h_{\mathbf{Y}}(\mathbf{y}) = a + bh^p$ *for the order* $p = 4$. (b) *Corresponding CPU time, relative to* 10,000 *iterations of the MCMC.* (c) *Posterior distribution of* λ *for the RK4 solver,* $p = 4$*, for step sizes and* $h = 0.00625$ *(blue histogram),* $h = 0.1$ *(green histogram), and exact posterior (black density). The former takes* 36 *minutes and the latter* 2.5.

latter is one order of magnitude larger than the former). This has to be linked to the remark we have made above: the error induced by the numerical integration of the ODE should not only be considered by itself, but also in light of the observation noise. When $\sigma = 30$, the step size h^* such that for any $h \leq h^*$ all the models $\mathcal{M}^h$ are equivalent on the Jeffreys scale is much higher, involving even larger computational time savings.

• Choosing a numerical solver with global error 4 or more (RK4, Adams type, etc.) will lead to a BF that is practically 1, within a range of step sizes h, due to the flat region in the polynomial $a + bh^p$; see Figure 2. This region depends not only on the solver, but also on the model and observation error variance. However, the solver order ensures that such a range exists. This then translates to being able to choose a bigger h, within that flat step size range, that will lead to virtually the same posterior distributions than using a far smaller step size, since the BF is basically 1 (see Figures 2, 4(a), and 6).

Table 1
Comparison of exact and estimated marginals for the Runge–Kutta method of order 4.

σ	$P_{\mathbf{Y}}(\mathbf{y})$	$\hat{P}_{\mathbf{Y}}(\mathbf{y})$
1	$1.854 \cdot 10^{-18}$	$1.862 \cdot 10^{-18}$
30	$1.638 \cdot 10^{-60}$	$1.699 \cdot 10^{-60}$

5.2. A diabetes minimal model. We now illustrate our results on a real data set and a more complex model for an oral glucose tolerance test (OGTT). After briefly describing the experiment and the model, we present our results.

An OGTT is performed for diagnosis of diabetes, metabolic syndrome, and other conditions. After a night's sleep, fasting patients are measured for blood glucose and asked to drink a sugar concentrate. Blood glucose is then measured over two and sometimes three hours, depending on local practices. We are developing a minimal model for blood glucose-insulin interaction based on a two-compartment model: one simple transfer compartment of glucose in the digestive system, and one more complex compartment for blood glucose and interactions with insulin and other glucose restitution mechanisms. Here we present this model to show our methodology estimating one parameter only (namely, the insulin sensitivity). Since there is only one parameter involved we are able to find the marginal $P^h_{\mathbf{Y}}(\mathbf{Y})$ by numerical integration, for comparison purposes.

Let $G(t)$ be the patient blood glucose level at time t in mg/dL. Let $I(t)$ be blood insulin level at time t and $L(t)$ the "glucagon" levels, to promote liver glycogen glucose production, in arbitrary units. Let $D(t)$ be the digestive system "glucose level"; we take it as a compartment in which glucose is first stored (e.g., stomach and digestive tract) and in turn delivered into the blood stream (we state $D(t)$ in the same units as for $G(t)$, and therefore the mean life parameter θ_2 in (5.5) is the same as the one used in (5.2) below). Let also G_b be the glucose baseline ($= 80$ mg/dL, fixed). Our model is described by the following system of ODEs:

$$\frac{dG}{dt} = (L - I)\, G + \frac{D}{\theta_2}, \tag{5.2}$$

$$\frac{dI}{dt} = \theta_0 \left(\frac{G}{G_b} - 1 \right)^+ - \frac{I}{a}, \tag{5.3}$$

$$\frac{dL}{dt} = \theta_1 \left(1 - \frac{G}{G_b} \right)^+ - \frac{L}{b}, \tag{5.4}$$

$$\frac{dD}{dt} = -\frac{D}{\theta_2}, \tag{5.5}$$

where $(x)^+$ means the positive part of x, i.e., $(x)^+ = x1(x > 0)$.

A brief explanation of the model follows. When glucose goes above the normal threshold G_b, insulin is produced, i.e., its derivative increases; see (5.3). This, in turn, acts on blood glucose to decrease its concentration; a mass-action-type term is introduced in (5.2) to decrease the derivative of G. L is an abstract term related to the glucose recovery system. When glucose $G(t)$ goes below the normal threshold (G_b, which we set to 80), L increases (see (5.4)) to increase the derivative of $G(t)$ (thus eventually increasing the glucose); see (5.2). Finally,

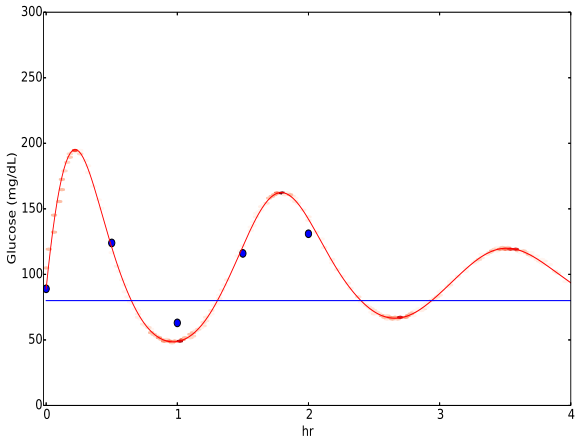


Figure 5. *OGTT test performed on an obese male adult, with glucose measurements taken every* 30 *minutes up to* 2 *hours (dots). Note the oscillating nature of the data, typically indicating a not well-controlled insulin-glucose system. The MAP model is shown in red, along with draws from the posterior predictive distribution shown in the shaded areas.*

$D(t)$ represents the glucose in the digestive compartment that will be transferred to the blood stream; see (5.5) and (5.2). We analyze data from an OGTT conducted in an obese male adult patient with a suspected metabolic syndrome condition; the corresponding data are plotted in Figure 5. All parameters are positive.

The principal aim in OGTT studies is the patient's insulin sensitivity, which in our model is θ_0, and this is the parameter to be inferred from the data. By experimenting with several data sets, reasonable behavior is obtained with $\theta_1 = 26.6$. The insulin and glucagon clearance rates are set to $a = 1, b = 2$, while the glucose transfer mean life $\theta_2 = 0.2$h (Anderwald et al. (2011)). The initial value $G(0)$ is set to d_0, the recorded fasting glucose. Homeostasis is assumed after a night's sleep, and therefore we set $I(0) = 0$ and $L(0) = 0$. The 75gr sugar concentrate translates into a maximum of $D(0) = 250\frac{mg}{dL}$, which is transferred into $G(t)$. This is an unusual experimental data set in which glucose was measured every 30 minutes up to 2 hours. Commonly, glucose is measured at arrival, at the hour, and at 2 hours only.

Our Bayesian inference is performed as follows. We have observations $d_1, d_2, \ldots, d_n$ for

$$d_i = G(t_i) + e_i, \quad \text{where} \quad e_i \sim N(0, \sigma^2),$$

and $G(0) = d_0$ the initial condition; we fix the measurement error to $\sigma = 5$ (the observation functional is therefore $f(X) = X_1$). From this a likelihood is constructed. The error here is due to the glucose measurement process and added error to account for model discrepancies, and a $\sigma = 5$ produces reasonable fit in most cases. Since θ_0 is positive, a Gamma prior distribution is assumed for parameter θ_0, with shape parameter $= 5$ and rate $= 2/5$, thus with mean $= 2$, apparently the value for a normal person (established from a small data set of OGTTs performed on healthy individuals). Using an RK4 solver with varying step size, we perform an MCMC for these parameters using the t-walk (Christen and Fox (2010)).

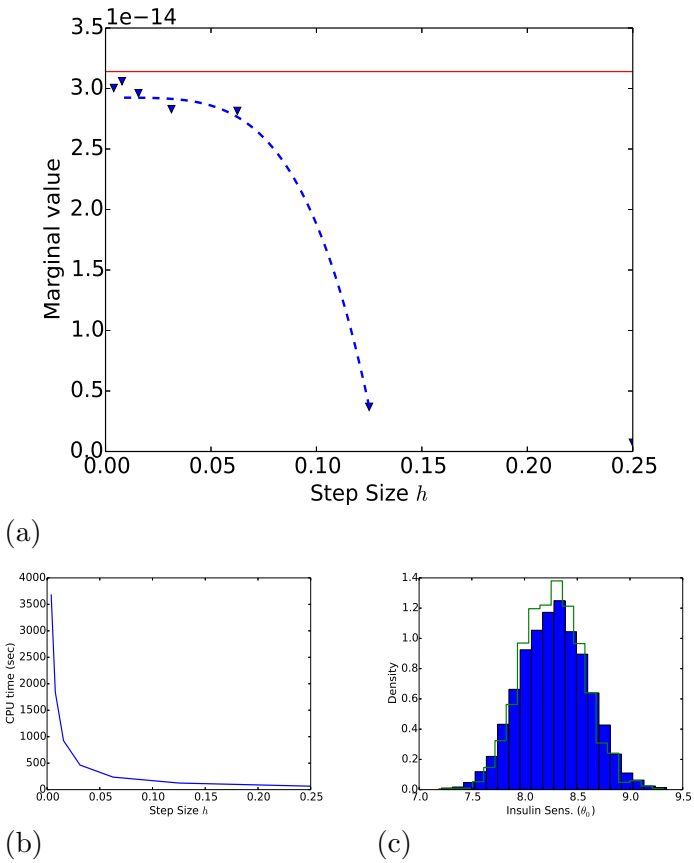


Figure 6. *BFs study for the diabetes minimal model using an order* 4 *Runge–Kutta solver.* (a) *The horizontal line (red) is the numerical integration approximation of* $P^h_{\mathbf{Y}}(\mathbf{Y})$ *using step size* $0.25 \cdot 2^{-7}$ *(smallest step size used), while the triangles are Monte Carlo estimates performed as in* (4.2)*; these seem to slightly underestimate the former. The dashed line (blue) is a regression model* $a + bh^4$ *estimate using step sizes from* $0.25 \cdot 2^{-1}$ *to* $0.25 \cdot 2^{-4}$ *only.* (b) *CPU time for various solver step sizes considering* 10,000 *iterations of the MCMC.* (c) *Comparison of the resulting posterior with step size* $0.25 \cdot 2^{-7}$ *(blue histogram) and* $0.25 \cdot 2^{-3}$ *(green histogram) showing basically no difference and resulting in a near* 90% *reduction in CPU evaluation time.*

To make the solver evaluation time steps (in hours) include the observation times, we take a rough time step of 15 minutes (= 0.25h) and divide it into finer time steps defined as $h^{(k)} = 0.25 \cdot 2^{-k}$h. Our experiments included $k = 0, \ldots, 7$, as seen in Figure 6(a). We only use an RK4 solver, resulting in the 4th grade polynomial regression and a flat section already at $h^{(3)} = 0.25 \cdot 2^{-3}\text{h} = 1.875\text{min}$. If compared with our minimum time step of $h^{(7)} = 0.25 \cdot 2^{-7}\text{h} = 7\text{s}$, the resulting CPU time of the MCMC is more than 90% larger. However, the resulting posterior distributions for $h^{(3)}$ and $h^{(7)}$ are basically identical (see Figure 6(c)). The estimated marginal is $P^{h^{(3)}}_{\mathbf{Y}}(\mathbf{y}) \approx 3 \cdot 10^{-14}$, calculated using the MCMC samples and (4.2), while $P^{h^{(7)}}_{\mathbf{Y}}(\mathbf{y}) \approx 3.2 \cdot 10^{-14}$, calculated using numerical integration (red line in Figure 6(a)).

6. Discussion. We build upon some theoretical aspects of the Bayesian analysis of ODE systems. As opposed to more standard (Bayesian) statistical analyses, inverse problems present the added difficulty that the regressor function (i.e., the forward map) is not analytically tractable and numerical approximations need to be used. In general, the replacement of the theoretical (unavailable) solution of the differential system by a numerical approximation is ignored, and the solver is used as a black box. Recently, however, research has been directed at trying to quantify the consequences of such an approximation, commonly by comparing expected values of the resulting posterior distributions, like the exact vs. the numerical posterior means.

In this paper we adopt a different approach, basing our comparison on the use of BFs, which is the natural tool for comparing models in a Bayesian context. There are still some particular issues to be solved when applying our results to general inverse problems like estimating the marginals in a multidimensional parameter problem and analyzing stiff problems where a multistep method would need to be used. However, we may highlight the following remarks.

First, we contribute to the intuitive idea that the ODE solver approximation error should be put in the perspective of the observational error. BFs, and the Bayesian model comparison machinery, may be used as an appropriate measure of solver accuracy, precisely from the perspective of the observational error considered in the model. Theorem 2 establishes a consistency between the ODE solver order and the BF of the exact vs. the numerical posterior distribution. As a consequence, the numerical solver used may be viewed from this perspective and not solely as a black box number crunching routine. As far as the main aim is to make inference on parameters, there is no need to use the highest precision if the data are contaminated by a nonnegligible quantity of noise. In a domain where the computational time is important, we have proved that considerable CPU time savings may be obtained only by using a reasonable step size in the solver.

Second, we show how the BF may be approximated even in this scenario where the exact model is not available. This result is of particular interest, since it allows us to compare the accuracy of our approximate posterior without being able to work on the theoretical model directly. The computation of marginal likelihoods is an important topic in the Bayesian literature. In this paper, we propose the use of the Gelfand and Dey estimator, which has the great advantage of not requiring any additional numerical evaluation of the differential system after the MCMC was performed. However, we are aware that the Gelfand and Dey estimator may become highly unstable as the dimension of the parameters increases. The use of a KDE weighting function in (4.2) may help to stabilize the estimate but is not a universal solution. If the dimension of the parameters increases, other strategies should be considered (Valpine (2008)), still keeping in mind that any additional numerical evaluations of the ODE system may have considerable computational costs. Our results would also need to be stated for multiple dimension observation functions f; we leave these considerations for future research.

Acknowledgments. We thank Dr. Silvia Quintana for kindly providing the OGTT data, for example, in section 5.2. We also thank the two reviewers and the AE comments and corrections, which greatly improved the presentation of the paper.

REFERENCES

C. Anderwald, A. Gastaldelli, A. Tura, M. Krebs, M. Promintzer-Schifferl, A. Kautzky-Willer, M. Stadler, R. DeFronzo, G. Pacini, and M. Bischof (2011), *Mechanism and effects of glucose absorption during an oral glucose tolerance test among females and males*, J. Clinical Endocrinol. Metabolism, 9, pp. 515–524.

A. Arnold, D. Calvetti, and E. Somersalo (2013), *Linear multistep methods, particle filtering and sequential Monte Carlo*, Inverse Problems, 29, 085007.

A. Arnold, D. Calvetti, and E. Somersalo (2014), *Parameter estimation for stiff deterministic dynamical systems via ensemble Kalman filter*, Inverse Problems, 30, 105008.

Y. Atchadé and J. Rosenthal (2005), *On adaptive Markov chain Monte Carlo algorithm*, Bernoulli, 11, pp. 815–828.

C. Cai, A. Mohammad-Djafari, S. Legoupil, and T. Rodet (2011), *Bayesian data fusion and inversion in x-ray multi-energy computed tomography*, in Proceedings of the 18th IEEE International Conference on Image Processing, ICIP, pp. 1377–1380.

D. Calvetti, R. K. Dash, E. Somersalo, and M. E. Cabrera (2006), *Local regularization method applied to estimating oxygen consumption during muscle activities*, Inverse Problems, 22, pp. 229–243.

Y. Cao and L. Petzold (2004), *A posteriori error estimation and global error control for ordinary differential equations by the adjoint method*, SIAM J. Sci. Comput., 26, pp. 359–374, doi:10.1137/S1064827503420969.

M. Capistrán, J. Christen, and Velasco-Hernández (2012), *Towards uncertainty quantification and inference in the stochastic SIR epidemic model*, Math. Biosci., 240, pp. 250–259.

Z. Chama, B. Mansouri, M. Anani, and A. Mohammad-Djafari (2012), *Image recovery from Fourier domain measurements via classification using Bayesian approach and total variation regularization*, AEU - Internat. J. Electron. Commun., 66, pp. 897–902.

J. Christen and C. Fox (2010), *A general purpose sampling algorithm for continuous distributions (the t-walk)*, Bayesian Anal., 5, pp. 263–282.

S. L. Cotter, M. Dashti, and A. M. Stuart (2010), *Approximation of Bayesian inverse problems for PDEs*, SIAM J. Numer. Anal., 48, pp. 322–345, doi:10.1137/090770734.

T. Cui, C. Fox, and M. J. O'Sullivan (2011), *Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm*, Water Resources Res., 47, W10521 2011.

S. Donnet and A. Samson (2007), *Estimation of parameters in missing data models defined by differential equations*, J. Statist. Plan. Inference, 137, pp. 2815–2831.

M. D. Fall, E. Barat, C. Comtat, T. Dautremer, T. Montagu, and A. Mohammad-Djafari (2011), *A discrete-continuous Bayesian model for emission tomography*, in Proceedings of the 18th IEEE International Conference on Image Processing, ICIP, pp. 1373–1376.

U. Foryś and A. Marciniak-Czochra (2003), *Logistic equations in tumour growth modelling*, Int. J. Appl. Math. Comput. Sci., 13, pp. 317–325.

C. Fox, M. Palm, and G. Nicholls (1999), *Efficient, exact PDE solutions for MCMC*, Proc. SPIE, 3816, pp. 23–30.

A. E. Gelfand and D. K. Dey (1994), *Bayesian model choice: Asymptotics and exact calculations*, J. Roy. Statist. Soc. Ser. B, 56, pp. 501–514.

R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna (2007), *Universally sloppy parameter sensitivities in systems biology models*, PLoS Comput. Biol., 3, e189.

H. Haario, E. Saksman, and J. Tamminen (1998), *An adaptive metropolis algorithm*, Bernoulli, 7, pp. 223–242.

M. L. Hazelton (2010), *Bayesian inference for network-based models with a linear inverse structure*, Transportation Res. Part B: Methodological, 44, pp. 674–685.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), *Bayesian model averaging: A tutorial*, Statist. Sci., 14, pp. 382–401.

A. Iserles (1996), *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, New York.

H. Jeffreys (1961), *Theory of Probability*, 3rd ed., Oxford University Press, Oxford.

J. P. Kaipio and C. Fox (2011), *The Bayesian framework for inverse problems in heat transfer*, Heat Transfer Engrg., 32, pp. 718–753.

J. P. Kaipio and E. Somersalo (2005), *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer, New York.

R. E. Kass and A. E. Raftery (1995), *Bayes factors*, J. Amer. Statist. Assoc., 90, pp. 773–795.

A. Keats, E. Yee, and F. Lien (2010), *Information-driven receptor placement for contaminant source determination*, Environmental Model. Software, 25, pp. 1000–1013.

V. Kolehmainen, A. Vanne, S. Siltanen, S. Järvenpää, J. P. Kaipio, M. Lassas, and M. Kalke (2007), *Bayesian inversion method for 3D dental x-ray imaging*, Elektrotech. Inform., 124, pp. 248–253.

S. Kozawa, T. Takenouchi, and K. Ikeda (2012), *Subsurface imaging for anti-personal mine detection by Bayesian super-resolution with a smooth-gap prior*, Artificial Life Robotics, 16, pp. 478–481.

J. D. Lambert (1991), *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem*, John Wiley & Sons, Chichester, UK.

J. Lang and J. G. Verwer (2007), *On global error estimation and control for initial value problems*, SIAM J. Sci. Comput., 29, pp. 1460–1475, doi:10.1137/050646950.

A. Mohammad-Djafari (2006), *Bayesian inference for inverse problems in signal and image processing and applications*, Int. J. Imaging Systems Technol., 16, pp. 209–214.

A. Nissinen, V. P. Kolehmainen, and J. P. Kaipio (2011), *Compensation of modelling errors due to unknown domain boundary in electrical impedance tomography*, IEEE Trans. Medical Imaging, 30, pp. 231–242.

A. Quarteroni, R. Sacco, and F. Saleri (2007), *Numerical Mathematics*, Texts Appl. Math. 37, Springer, Berlin.

K. Radhakrishnan and A. Hindmarsh (1993), *Description and Use of LSODE, the Livermore Solver for Ordinary Differential Equations*, Tech. report, Lawrence Livermore National Laboratory.

C. Robert and G. Casella (2004), *Monte Carlo Statistical Methods*, Springer, New York.

C. Schwab and A. M. Stuart (2012), *Sparse deterministic approximation of Bayesian inverse problems*, Inverse Problems, 28, 045003.

E. Somersalo, A. Voutilainen, and J. P. Kaipio (2003), *Non-stationary magnetoencephalography by Bayesian filtering of dipole models*, Inverse Problems, 19, pp. 1047–1063.

E. Süli and D. F. Mayers (2003), *An Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, UK.

P. D. Valpine, (2008). *Improved estimation of normalizing constants from Markov chain Monte Carlo output*, J. Comput. Graph. Statist., 17, pp. 333–351.

A. Vehtari and J. Lampinen (2000), *Bayesian MLP neural networks for image analysis*, Pattern Recognition Lett., 21, pp. 1183–1191.

J. Wan and N. Zabaras (2011), *A Bayesian approach to multiscale inverse problems using the sequential Monte Carlo method*, Inverse Problems, 27, 105004.

D. Watzenig and C. Fox (2009), *A review of statistical modelling and inference for electrical capacitance tomography*, Measurement Sci. Technol., 20 (2009), 052002.

A. D. Woodbury, (2011), *Minimum relative entropy, Bayes and Kapur*, Geophys. J. Internat. 185, pp. 181–189.

H. Xue, H. Miao, and H. Wu (2010), *Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error*, Ann. Statist., 38, pp. 2351–2387.

S. Zhu, P. You, H. Wang, X. Li, and A. Mohammad-Djafari (2011), *Recognition-oriented Bayesian SAR imaging*, in Proceedings of the 2011 3rd International Asia-Pacific Conference on Synthetic Aperture Radar, APSAR, 2011, pp. 153–156.