UNCECOMP 2015
1st ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Crete Island, Greece, 25–27 May 2015

VARIATIONAL REFORMULATION OF BAYESIAN INVERSE PROBLEMS

Panagiotis Tsilifis[1], Ilias Bilionis[2†], Ioannis Katsounaros[3a,3b,3c] and Nicholas Zabaras[4]

[1]Department of Mathematics, University of Southern California, Los Angeles, CA 90089-2532, USA
e-mail: tsilifis@usc.edu

[2] School of Mechanical Engineering, Purdue University, 585 Purdue Mall, West Lafayette, IN 47906-2088, USA
e-mail: ibilion@purdue.edu

[3a]Department of Chemistry, University of Illinois at Urbana-Champaign, S. Mathews Ave., Urbana, IL 61801, USA
[3b]Materials Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Lemont, IL 60439, USA
[3c]Leiden Institute of Chemistry, Leiden University, Einsteinweg 55, P.O. Box 9502, 2300 RA Leiden, The Netherlands
e-mail: katsounaros@anl.gov

[4]Warwick Centre for Predictive Modelling, The University of Warwick, Coventry, CV4 7AL, UK
e-mail: nzabaras@gmail.com

Keywords: Variational inference, bayesian, fast approximation, kullback-leibler

Abstract. *The classical approach to inverse problems is based on the optimization of a misfit function. Despite its computational appeal, such an approach suffers from many shortcomings, e.g., non-uniqueness of solutions, modeling prior knowledge, etc. The Bayesian formalism to inverse problems avoids most of the difficulties encountered by the optimization approach, albeit at an increased computational cost. In this work, we use information theoretic arguments to cast the Bayesian inference problem in terms of an optimization problem. The resulting scheme combines the theoretical soundness of fully Bayesian inference with the computational efficiency of a simple optimization.*

†Corresponding author.

1 INTRODUCTION

As we are approaching the era of exascale computing, we encounter more and more complex physical models. These complex models have many unknown parameters that need to be inferred from experimental measurements. That is, inverse problems are becoming an integral part of every scientific discipline that attempts to combine computational models with data. These include climate modeling [4], numerical weather prediction [18, 25], subsurface hydrology and geology [15], and many more.

It can be argued that the "right" way to pose an inverse problem is to follow the Bayesian formalism [28, 17]. It is the "right" way because it deals with three well-known difficulties of inverse problems: non-uniqueness issues, modeling prior knowledge, and estimating experimental noise. The Bayesian solution of an inverse problem is summarized neatly by the *posterior* probability density of the quantity of interest. In turn, the posterior can only be explored numerically by Monte Carlo (MC) methods, the most successful of which is Markov Chain Monte Carlo (MCMC) [24, 14]. Because of the need to repeatedly evaluate the underlying physical model, MCMC is computationally impractical for all but the simplest cases. Therefore, we need methods that approximate the posterior in a computationally efficient manner.

One way to come up with a computationally attractive approximation is to replace the physical model with a cheap-to-evaluate surrogate [20, 22]. In this way, MCMC becomes feasible again. However, there is no free lunch: firstly, things become complicated when the surrogate is inaccurate, and, secondly, constructing accurate surrogates is exponentially hard as the number of parameters increase [3]. Because of these difficulties, in this work, we attempt to develop a surrogate-free methodology.

Perhaps the simplest idea is to approximate the posterior with a delta function centered about its maximum. The maximum of the posterior is known as the maximum a posteriori (MAP) estimate of the parameters. The MAP approach is justified if the posterior is sharply picked around a unique maximum. It requires the solution of an optimization problem. The objective function of this optimization has two parts: a misfit and a regularization part that arise from the likelihood and the prior, respectively. The MAP estimate is commonly used in numerical weather prediction problems [25] as well as in seismic inversion [8].

The Laplace approximation represents the posterior by a Gaussian density with a mean given by the MAP and a covariance matrix given by the negative inverse Hessian of the logarithm of the posterior. The Laplace approximation is justified when the posterior has a unique maximum and is shaped, more or less, like a Gaussian around it. As before, the identification of the MAP requires the solution of an optimization problem. The required Hessian information may be estimated numerically either by automatic differentiation methods [12] or by developing the adjoint equations of the underlying physical model [26].

Variational inference (VI) [10, 7] is a class of techniques in Bayesian statistics targeted toward the construction of approximate posteriors. One proceeds by posing a variational problem whose solution over a family of probability densities approximates the posterior. VI techniques have been proved quite effective for a wide class of inference problems. However, in their archetypal form, they are not directly applicable to inverse problems. This is due to the, typically, non-analytic nature of the underlying physical models. Fortunately, the recent developments in non-parametric VI by [11] can come to rescue. This is the approach we follow in this work. In non-parametric VI, the family of probability densities that serve as candidate posteriors is the family of mixtures of Gaussians with a fixed number of components [23]. Since a mixture of Gaussians with an adequate number of components can represent any probability density,

this approximating family is sufficiently large. The approximate posterior is constructed by minimizing the information loss between the true posterior and the approximate one over the candidate family. This is achieved by solving a series of optimization problems [7].

The outline of the paper is as follows. We start Sec. 2 with a generic discussion of the Bayesian formulation of inverse problems. In Sec. 2.1 we present the core ideas of VI and in Sec. 2.2 we show how one can develop approximation schemes. We validate the proposed methodology numerically by solving two inverse problems: the estimation of the kinetic parameters of a catalysis system (Sec. 3.1) and the identification of the source of contamination based on scarce concentration measurements (Sec. 3.2). Finally, in A we provide all the technical details that are required for the implementation of the proposed methodology.

2 METHODOLOGY

A forward model associated with a physical phenomenon can be thought of as a function $\mathbf{f} : \mathbb{R}^{d_\xi} \to \mathbb{R}^{d_y}$, that connects some unknown parameters $\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$ to some observed quantities $\mathbf{y} \in \mathbb{R}^{d_y}$. This connection is defined indirectly via a *likelihood* function:

$$p(\mathbf{y}|\boldsymbol{\xi}, \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{f}(\boldsymbol{\xi}), \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$ denotes the parameters that control the measurement noise and/or the model discrepancy. Notice how in Eq. (1) the observations, $\mathbf{y}$, are explicitly connected to the parameters, $\boldsymbol{\xi}$, through the model, $\mathbf{f}(\boldsymbol{\xi})$. A typical example of a likelihood function is the *isotropic Gaussian likelihood*:

$$p(\mathbf{y}|\boldsymbol{\xi}, \theta) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}(\boldsymbol{\xi}), e^{2\theta}\mathbf{I}\right), \tag{2}$$

where θ is a real number, and $\mathbf{I}$ is the unit matrix with the same dimensions as $\mathbf{y}$. The exponential of the parameter,

$$\sigma = e^{\theta}, \tag{3}$$

can be thought of as the standard deviation of the measurement noise. The usual parameterization of the isotropic Gaussian likelihood uses σ instead of θ. We do not follow the usual approach because in our numerical examples, it is preferable to work with real rather than positive numbers.

Both $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are characterized by *prior* probability densities, $p(\boldsymbol{\xi})$ and $p(\boldsymbol{\theta})$, respectively. These describe our state of knowledge, prior to observing $\mathbf{y}$. As soon as $\mathbf{y}$ is observed, our updated state of knowledge is captured by the *posterior* distribution:

$$p(\boldsymbol{\xi}, \boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\xi}, \boldsymbol{\theta})p(\boldsymbol{\xi})p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{4}$$

where the normalization constant $p(\mathbf{y})$,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\xi}, \boldsymbol{\theta})p(\boldsymbol{\xi})p(\boldsymbol{\theta})d\boldsymbol{\xi}d\boldsymbol{\theta}, \tag{5}$$

is known as the *evidence*. Eq. (4) is the Bayesian solution to the inverse problem. Notice that it is a probability density over the joint space of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. This is to be contrasted with the classical approaches to inverse problems whose solutions result in point estimates of the unknown variables. The mass of this probability density corresponds to our inability to fully resolve the values of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ due to insufficient experimental information.

Notice that by writing $\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\theta}) \in \mathbb{R}^d$,$d = d_\xi + d_\omega$, $p(\boldsymbol{\omega}) = p(\boldsymbol{\xi})p(\boldsymbol{\theta})$, and $p(\mathbf{y}|\boldsymbol{\omega}) = p(\mathbf{y}|\boldsymbol{\xi}, \boldsymbol{\theta})$, we may simplify the notation of Eq. (4) to

$$p(\boldsymbol{\omega}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\omega})p(\boldsymbol{\omega})}{p(\mathbf{y})}. \tag{6}$$

This is the notation we follow through out this work. The goal of the rest of the paper is to construct an algorithmic framework for the efficient approximation of the posterior of Eq. (6).

2.1 Variational inference

Consider a family $\mathcal{Q}$ of probability densities over $\boldsymbol{\omega}$. Our objective is to choose a $q(\boldsymbol{\omega}) \in \mathcal{Q}$ that is as "close" as possible to the posterior $p(\boldsymbol{\omega}|\mathbf{y})$ of Eq. (6). This "closeness" is measured by the Kullback-Leibler (KL) divergence [21],

$$\mathrm{KL}\left[q(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega}|\mathbf{y})\right] = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{\omega})}{p(\boldsymbol{\omega}|\mathbf{y})}\right], \tag{7}$$

where $\mathbb{E}_q[\cdot]$ denotes the expectation with respect to $q(\boldsymbol{\omega})$. Intuitively, the KL divergence can be thought of as the *information loss* we experience when we approximate the posterior $p(\boldsymbol{\omega}|\mathbf{y})$ with the probability density $q(\boldsymbol{\omega})$. It is easy to show that

$$\mathrm{KL}\left[q(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega}|\mathbf{y})\right] \geq 0, \tag{8}$$

and that the equality holds if and only if $q(\boldsymbol{\omega}) = p(\boldsymbol{\omega}|\mathbf{y})$. Therefore, if the posterior is in $\mathcal{Q}$, then minimizing Eq. (7) over $q(\omega) \in \mathcal{Q}$ will give an exact answer. For an arbitrary choice of $\mathcal{Q}$, we postulate that minimizing Eq. (7) yields a good approximation to the posterior.

Unfortunately, calculation of Eq. (7) requires knowledge of the posterior. This means that Eq. (7) cannot be used directly in an optimization scheme. In order to proceed, we need an objective function that does not depend explicitly on the posterior. To this end, notice that the evidence may be expressed as:

$$\log p(\mathbf{y}) = \mathcal{F}[q] + \mathrm{KL}\left[q(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega}|\mathbf{y})\right], \tag{9}$$

where

$$\mathcal{F}[q] = \mathbb{E}_q\left[\log \frac{p(\mathbf{y}, \boldsymbol{\omega})}{q(\boldsymbol{\omega})}\right] = \mathcal{H}[q] + \mathbb{E}_q\left[\log p(\mathbf{y}, \boldsymbol{\omega})\right], \tag{10}$$

with

$$p(\mathbf{y}, \boldsymbol{\omega}) = p(\mathbf{y}|\boldsymbol{\omega})p(\boldsymbol{\omega}) \tag{11}$$

being the joint probability density of $\mathbf{y}$ and $\boldsymbol{\omega}$, and

$$\mathcal{H}[q] = -\mathbb{E}_q\left[\log q(\boldsymbol{\omega})\right] \tag{12}$$

being the *entropy* of $q(\boldsymbol{\omega})$. Since the KL divergence is non-negative (Eq. (8)), we have from Eq. (9) that

$$\mathcal{F}[q] \leq \log p(\mathbf{y}). \tag{13}$$

The functional $\mathcal{F}[q]$ is generally known as the *evidence lower bound* (ELBO).

We see, that maximizing Eq. (10) is equivalent to minimizing Eq. (7). In addition, Eq. (10) does not depend on the posterior in an explicit manner. This brings us to the variational principle used through out this work: The "best" approximation to the posterior of Eq. (6) over the family of probability densities $\mathcal{Q}$ is the solution to the following optimization problem:

$$q^*(\boldsymbol{\omega}) = \arg\max_q \mathcal{F}[q]. \tag{14}$$

2.2 Developing approximation schemes

The main difficulty involved in the solution Eq. (14) is the evaluation of expectations over $q(\boldsymbol{\omega})$. In principle, one can approximate these expectations via a Monte Carlo procedure and, then, use a variant of the Robbins-Monro algorithm [27]. Such an approach yields a stochastic optimization scheme in the spirit of [5], and [2]. Whether or not such a scheme is more efficient than MCMC sampling of the posterior is beyond the scope of this work. Here, we follow the approach outlined by [11]. In particular, we will derive analytical approximations of $\mathcal{F}[q]$ for the special case in which $\mathcal{Q}$ is the family of Gaussian mixtures.

The family of Gaussian mixtures with L components is the family $\mathcal{Q}_L$ of probability densities of the form:

$$q(\boldsymbol{\omega}) = \sum_{i=1}^{L} w_i \mathcal{N}(\boldsymbol{\omega}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{15}$$

where w_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the responsibility, mean, and covariance matrix, respectively, of the i-th component of the mixture. The responsibilities w_i are non-negative and they sum to one while the covariance matrices $\boldsymbol{\Sigma}_i$ are positive definite. When we work with $\mathcal{Q}_L$, the generic variational problem of Eq. (6) is equivalent to optimization with respect to all the w_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. In what follows, we replace the ELBO, $\mathcal{F}[q]$ of Eq. (10), with a series of analytic approximations that exploit the properties of $\mathcal{Q}_L$, and, finally, we derive a three-step optimization scheme that yields a local maximum of the approximate ELBO.

We start with an approximation to the entropy $\mathcal{H}[q]$ (see Eq. (12)) of a Gaussian mixture Eq. (15). There are basically two kinds of approximations that may be derived: 1) using Jensen's inequality yields a lower bound to $\mathcal{H}[q]$ built out of convolutions of Gaussians (see [11] and [16]), and 2) employing a Taylor expansion of $\log q(\boldsymbol{\omega})$ about each $\boldsymbol{\mu}_i$ and evaluating the expectation over $\mathcal{N}(\boldsymbol{\omega}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ (see [16]). We have experimented with both approximations to the entropy without observing any significant differences in the numerical results. Therefore, we opt for the former one since it has a very simple analytical form. An application of Jensen's inequality followed by well-known results about the convolution of two Gaussians yields

$$\mathcal{H}[q] \geq \mathcal{H}_0[q], \tag{16}$$

where

$$\mathcal{H}_0[q] = -\sum_{i=1}^{L} w_i \ln q_i, \tag{17}$$

with

$$q_i = \sum_{j=1}^{L} w_j \mathcal{N}(\boldsymbol{\mu}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j). \tag{18}$$

The idea is to simply replace $\mathcal{H}[q]$ in Eq. (10) with $\mathcal{H}_0[q]$ of Eq. (17). This results in a lower bound to the ELBO $\mathcal{F}[q]$.

Now, we turn our attention to the second term of Eq. (10). For convenience, let us write it as:

$$\mathcal{L}[q] = \mathbb{E}_q \left[\log p(\mathbf{y}, \boldsymbol{\omega})\right]. \tag{19}$$

Notice that $\mathcal{L}[q]$ may be expanded as:

$$\mathcal{L}[q] = \sum_{i=1}^{L} w_i \mathbb{E}_{\mathcal{N}(\boldsymbol{\omega}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \left[\log p(\mathbf{y}, \boldsymbol{\omega})\right], \tag{20}$$

and that each expectation term can be approximated by taking the Taylor expansion of $\log p(\mathbf{y}, \boldsymbol{\omega})$ about $\boldsymbol{\omega} = \boldsymbol{\mu}_i$:

$$\begin{aligned} \log p(\mathbf{y}, \boldsymbol{\omega}) \quad \approx \quad & \log p(\mathbf{y}, \boldsymbol{\omega} = \boldsymbol{\mu}_i) + \nabla_{\boldsymbol{\omega}} \log p(\mathbf{y}, \boldsymbol{\omega} = \boldsymbol{\mu}_i) \, (\boldsymbol{\omega} - \boldsymbol{\mu}_i) \\ & + \tfrac{1}{2} (\boldsymbol{\omega} - \boldsymbol{\mu}_i)^T \nabla^2_{\boldsymbol{\omega}} \log p(\mathbf{y}, \boldsymbol{\omega} = \boldsymbol{\mu}_i) (\boldsymbol{\omega} - \boldsymbol{\mu}_i), \end{aligned} \tag{21}$$

where $\nabla_{\boldsymbol{\omega}}$ and $\nabla^2_{\boldsymbol{\omega}}$ stand for the Jacobian and the Hessian with respect to $\boldsymbol{\omega}$, respectively. Combining Eq. (20) with Eq. (21), we get the zero and second order Taylor approximation to $\mathcal{L}[q]$ of Eq. (19),

$$\mathcal{L}_0[q] = \sum_{i=1}^{L} w_i \log p(\mathbf{y}, \boldsymbol{\omega} = \boldsymbol{\mu}_i), \tag{22}$$

and

$$\mathcal{L}_2[q] = \mathcal{L}_0[q] + \frac{1}{2} \sum_{i=1}^{L} w_i \operatorname{Tr} \left[\boldsymbol{\Sigma}_i \nabla^2_{\boldsymbol{\omega}} \log p(\mathbf{y}, \boldsymbol{\omega} = \boldsymbol{\mu}_i) \right], \tag{23}$$

respectively.

Combining Eq. (17) with Eq. (22) or Eq. (23) we get an approximation to Eq. (10). In particular, we define:

$$\mathcal{F}_a[q] = \mathcal{H}_0[q] + \mathcal{L}_a[q], \tag{24}$$

where $a = 1$, or 2, selects the approximation to Eq. (19). From this point on, our goal is to derive an algorithm that converges to a local maximum of $\mathcal{F}_2[q]$.

Notice that $\mathcal{F}_2[q]$ requires the Hessian of $\log p(\mathbf{y}, \boldsymbol{\omega})$ at $\boldsymbol{\omega} = \boldsymbol{\mu}_i$. Therefore, optimizing it with respect to $\boldsymbol{\mu}_i$ would require third derivatives of $\log p(\mathbf{y}, \boldsymbol{\omega})$. This, in turn, implies the availability of third derivatives of the forward model $\mathbf{f}(\boldsymbol{\xi})$. Getting third derivatives of the forward model is impractical in almost all cases. In contrast, optimization of $\mathcal{F}_0[q]$ with respect to $\boldsymbol{\mu}_i$ requires only first derivatives of $\log p(\mathbf{y}, \boldsymbol{\omega})$, i.e., only first derivatives of the forward model $\mathbf{f}(\boldsymbol{\xi})$. In many inverse problems of interest, derivatives can be obtained at a minimum cost by making use of adjoint techniques (e.g., [9]). Therefore, optimization of $\mathcal{F}_{00}[q]$ with respect to $\boldsymbol{\mu}_i$ is computationally feasible.

The situation for $\boldsymbol{\Sigma}_i$ is quite different. Firstly, notice that $\mathcal{H}_0[q]$ increases logarithmically without bounds as a function of $|\boldsymbol{\Sigma}_i|$ and that the $\mathcal{L}_0[q]$ does not depend on $\boldsymbol{\Sigma}_i$ at all. We see that the lowest approximation that carries information about $\boldsymbol{\Sigma}_i$ is $\mathcal{F}_2[q]$. Looking at Eq. (23) we observe that this information is conveyed via the Hessian of $\log p(\mathbf{y}, \boldsymbol{\omega})$. This, in turn, requires the Hessian of the forward model $\mathbf{f}(\boldsymbol{\xi})$. The latter is a non-trivial task which is, however, feasible. In addition, notice that if $\boldsymbol{\Sigma}_i$ is restricted to be diagonal, then only the diagonal part of the Hessian of $\log p(\mathbf{y}, \boldsymbol{\omega})$ is required, thus, significantly reducing the memory requirements. The computation of $F_2[q]$ as well as of its gradients with respect to w_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ is discussed in A.

Taking the above into account, we propose an optimization scheme that alternates between optimizing $\{\boldsymbol{\mu}_i\}_{i=1}^L$, $\{\boldsymbol{\Sigma}_i\}_{i=1}^L$, and $\{w_i\}_{i=1}^L$. The algorithm is summarized in Algorithm 1. However, in order to avoid the use of third derivatives of the forward model we follow [11] in using $\mathcal{F}_0[q]$ as the objective function when optimizing for $\{\boldsymbol{\mu}_i\}$. Furthermore, we restrict our attention to diagonal covariance matrices,

$$\boldsymbol{\Sigma}_i = \operatorname{diag} \left(\sigma^2_{i1}, \ldots, \sigma^2_{id} \right), \tag{25}$$

since we do not want to deal with the issue of enforcing positive definiteness of the $\mathbf{\Sigma}_i$'s during their optimization. We use the L-BFGS-B algorithm of [6], which can perform bound constrained optimization. The bounds we use are problem-specific and are discussed in Sec. 3. In all numerical examples, we use the same convergence tolerance $\epsilon = 10^{-2}$.

Algorithm 1: Variational Inference

Input : Data $\mathbf{y}$, number of components L.
Initialize: $w_i = 1/N$, $\mathbf{\Sigma}_i = I$, and $\boldsymbol{\mu}_i$ randomly, for $i = 1, \ldots, L$.
repeat
  for $i = 1$ *to* L **do**
    $\{\boldsymbol{\mu}_i\} \leftarrow \arg\max_{\{\boldsymbol{\mu}_i\}} \mathcal{F}_0[q]$
    $\{w_i\} \leftarrow \arg\max_{\{w_i\}} \mathcal{F}_2[q]$
    $\{\mathbf{\Sigma}_i\} \leftarrow \arg\max_{\left\{\mathbf{\Sigma}_i = \mathrm{diag}\left(\sigma_{i1}^2, \ldots, \sigma_{id}^2\right)\right\}} \mathcal{F}_2[q]$
  end
until *change in* $\mathcal{F}_2[q]$ *is less than a tolerance* ϵ

3 EXAMPLES

In this section we present two numerical examples: 1) The problem of estimating rate constants in a catalysis system (Sec. 3.1), and 2) the problem of identifying the source of contamination in a two dimensional domain (Sec. 3.2). In both examples, we compare the approximate posterior to a MCMC [14] estimate. We used the Metropolis-Adjusted-Langevin-Algorithm (MALA) [1], since it can use the derivatives of the forward models to accelerate convergence. In all examples, the step size of the MALA proposal was $dt = 0.1$, the first $1,000$ steps were burned, and we observed the chain every 100 steps for a total of $100,000$ steps. We implented our methodology is Python. The code is freely available at `https://github.com/ebilionis/variational-reformulation-of-inverse-problems`.

3.1 Reaction kinetic model

We consider the problem of estimating kinetic parameters of multi-step chemical reactions that involve various intermediate or final products. In particular, we study the dynamical system that describes the catalytic conversion of nitrate (NO_3^-) to nitrogen (N_2) and other by-products by electrochemical means. The mechanism that is followed is complex and not well understood. In this work, we use the simplified model proposed by [19], which includes the production of nitrogen (N_2), ammonia (NH_3), and nitrous oxide (N_2O) as final products, as well as that of nitrite (NO_2^-) and an unknown species X as reaction intermediates (see Fig. 1).

The dynamical system associated with the reaction depicted in Fig. 1 is:

$$\begin{aligned}
\tfrac{d[\mathrm{NO_3^-}]}{dt} &= -k_1[\mathrm{NO_3^-}], \\
\tfrac{d[\mathrm{NO_2^-}]}{dt} &= k_1[\mathrm{NO_3^-}] - (k_2 + k_4 + k_5)[\mathrm{NO_2^-}], \\
\tfrac{d[\mathrm{X}]}{dt} &= k_2[\mathrm{NO_2^-}] - k_3[X], \\
\tfrac{d[\mathrm{N_2}]}{dt} &= k_3[\mathrm{X}], \\
\tfrac{d[\mathrm{NH_3}]}{dt} &= k_4[\mathrm{NO_2^-}], \\
\tfrac{d[\mathrm{N_2O}]}{dt} &= k_5[\mathrm{NO_2^-}],
\end{aligned} \tag{26}$$

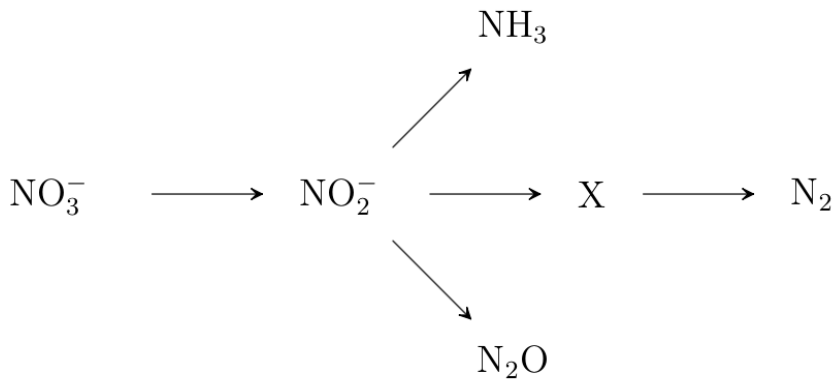

Figure 1: Reaction kinetic model: Simplified reaction scheme for the reduction of nitr4ate.

where $[\cdot]$ denotes the concentration of a quantity, and $k_i > 0$, $i = 1, ...5$ are the *kinetic rate constants*. Our goal is to estimate the kinetic rate constants based on the experimental measurements obtained by [19]. For completeness, these measurements are reproduced in Table 1. The initial conditions for Eq. (26) are given by the $t = 0$ row of the Table 1.

t (min)	$[NO_3^-]$	$[NO_2^-]$	[X]	$[N_2]$	$[NH_3]$	$[N_2O]$
0	500.00	0.00	-	0.00	0.00	0.00
30	250.95	107.32	-	18.51	3.33	4.98
60	123.66	132.33	-	74.85	7.34	20.14
90	84.47	98.81	-	166.19	13.14	42.10
120	30.24	38.74	-	249.78	19.54	55.98
150	27.94	10.42	-	292.32	24.07	60.65
180	13.54	6.11	-	309.50	27.26	62.54

Table 1: Reaction kinetic model: The table contains the experimental data used in the calibration process. The rows correspond to the time of each measurement and the columns to the concentrations measured in mmol $\cdot$ L^{-1}. The "-" symbols corresponds to lack of observations. See [19] for more details on the experiment. The $t = 0$ row provides the initial condition to (**??**). The observed data vector $\mathbf{y} \in \mathbb{R}^{30}$ is built by concatenating the scaled version of the rows $t = 30$ to $t = 180$ while skipping the row corresponding to X.

In order to avoid numerical instabilities, we work with a dimensionless version of Eq. (26). In particular, we define the scaled time:

$$\tau = \frac{t}{180 \text{ min}}, \tag{27}$$

the scaled concentrations:

$$u_i = \frac{[\text{Y}]}{500 \text{ mmol} \cdot \text{L}^{-1}}, \tag{28}$$

for $i = 1, 2, 3, 4, 5, 6, 7$, and $\text{Y} = \text{NO}_3^-, \text{NO}_2^-, \text{X}, \text{N}_2, \text{NH}_3, \text{N}_2\text{O}$, respectively, and the scaled kinetic rate constants:

$$\kappa_i = k_i \cdot 180 \text{ min}, \tag{29}$$

for $i = 1, \ldots, 5$. The dimensionless dynamical system associated with Eq. (26) is:

$$\begin{aligned} \dot{u}_1 &= -\kappa_1 u_1, \\ \dot{u}_2 &= \kappa_1 u_1 - (\kappa_2 + \kappa_4 + \kappa_5) u_2, \\ \dot{u}_3 &= \kappa_2 u_2 - \kappa_3 u_3, \\ \dot{u}_4 &= \kappa_3 u_3, \\ \dot{u}_5 &= \kappa_4 u_2, \\ \dot{u}_6 &= \kappa_5 u_2, \end{aligned} \tag{30}$$

where $\dot{u}$ denotes the differentiation of u with respect to the scaled time τ. The initial conditions for Eq. (30) are given by the scaled version of the $t = 0$ row of Table 1. We arrange the scaled version of the experimental data of Table 1 in a vector form, $\mathbf{y} \in \mathbb{R}^{d_y}$ with $d_y = 30$, by concatenating rows $t = 30, \ldots, 180$ of Table 1 while skipping the third column.

Table 2: Reaction kinetic model: The logarithm of the scaled kinetic rate constants, ξ_i (see Eq. (31), and the logarithm of the likelihood noise, θ (see Eq. (2)) as estimated by the variational approach with $L = 1$ and MCMC (MALA). The estimates correspond to the mean and the uncertainties to two times the standard deviation of each method.

Variable	VAR ($L = 1$)	MCMC (MALA)
ξ_1	1.359 ± 0.055	1.356 ± 0.072
ξ_2	1.657 ± 0.086	1.664 ± 0.142
ξ_3	1.347 ± 0.118	1.349 ± 0.215
ξ_4	-0.162 ± 0.167	-0.159 ± 0.230
ξ_5	-1.009 ± 0.368	-1.071 ± 0.513
θ	-3.840 ± 0.204	-3.757 ± 0.251

Since the kinetic rate constants, κ_i, of the scaled dynamical system of Eq. (30) are non-negative, it is problematic to attempt to approximate the posteriors associated with them with mixtures of Gaussians. For this reason, we work with the logarithms of the κ_i's,

$$\xi_i = \log \kappa_i, \tag{31}$$

for $i = 1, \ldots, 5$. We collectively denote those variables by $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_5)$. The prior probability density we assign to $\boldsymbol{\xi}$ is:

$$p(\boldsymbol{\xi}) = \prod_{i=1}^{5} p(\xi_i), \tag{32}$$

with

$$p(\xi_i) = \mathcal{N}(\xi_i | 0, 1), \tag{33}$$

for $i = 1, \ldots, 5$.

The forward model $\mathbf{f} : \mathbb{R}^5 \to \mathbb{R}^{30}$ associated with the experimental observations $\mathbf{y}$ is:

$$\mathbf{f}(\boldsymbol{\xi}) = \begin{pmatrix} u_1(t_2, \boldsymbol{\xi}), u_2(t_2, \boldsymbol{\xi}), u_4(t_2, \boldsymbol{\xi}) u_5(t_2, \boldsymbol{\xi}), u_6(t_2, \boldsymbol{\xi}), \\ \ldots \\ u_1(t_7, \boldsymbol{\xi}), u_2(t_7, \boldsymbol{\xi}), u_4(t_7, \boldsymbol{\xi}) u_5(t_7, \boldsymbol{\xi}), u_6(t_7, \boldsymbol{\xi}), \end{pmatrix}, \tag{34}$$

where $u_i(t, \boldsymbol{\xi})$ is the solution of Eq. (30) with the initial conditions specified by the scaled version of the $t = 0$ row of Table 1, and scaled kinetic rate constants, κ_i, given by inverting

Rate constant	Katsounaros (2012)	VAR Median	VAR 95% Interval
k_1	0.0216 ± 0.0014	0.0216	(0.0205, 0.0229)
k_2	0.0292 ± 0.0036	0.0291	(0.0269, 0.0316)
k_3	0.0219 ± 0.0044	0.0214	(0.0191, 0.0239)
k_4	0.0021 ± 0.0008	0.0020	(0.0014, 0.0030)
k_5	0.0048 ± 0.0008	0.0047	(0.0040, 0.0056)
σ	Not available	0.0215	(0.0176, 0.0262)

Table 3: Reaction kinetic model: The first five rows compare the median and 95% credible intervals of the kinetic rate constants estimated via the variational approach to those found in [19]. The units of the rates are in min^{-1}. The last line shows the median and 95% credible interval of the scaled measurement noise $\sigma = e^{\theta}$. This quantity is unit-less.

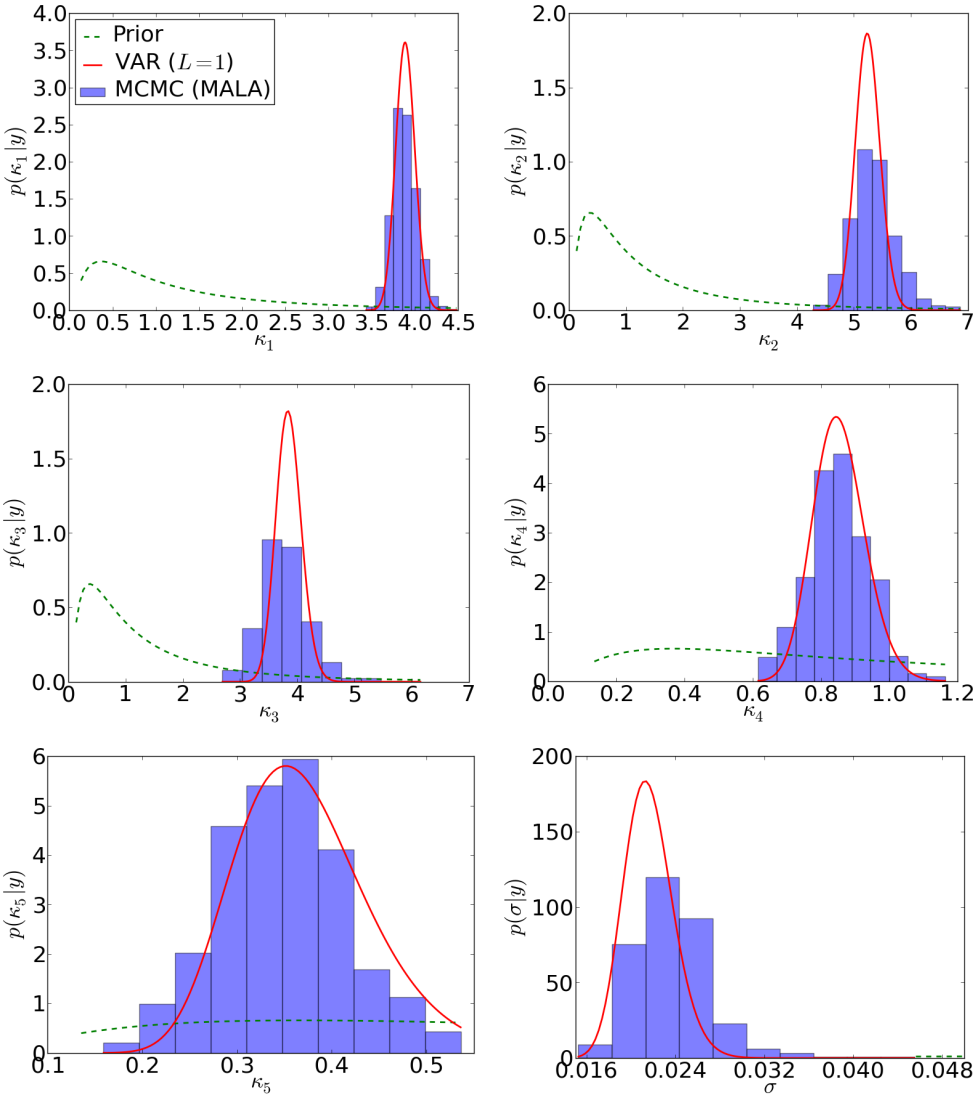

Figure 2: Reaction kinetic model: Comparison of the variational posterior (VAR ($L = 1$)) of the scaled kinetic rate constants κ_i as well as of the likelihood noise σ to the MCMC (MALA) histograms of the same quantity. The prior probability density of each quantity is shown as a dashed green line.

Eq. (31), i.e. $\kappa_i = e^{\xi_i}$. The derivatives of $\mathbf{f}(\boldsymbol{\xi})$ can be solving a series of dynamical systems forced by the solution of Eq. (30). This is discussed in A.5.

We use the isotropic Gaussian likelihood defined in Eq. (2). It is further discussed in A.4. The prior we assign to the parameter θ of the likelihood is:

$$p(\theta) = \mathcal{N}(\theta| -1, 1). \tag{35}$$

Since the noise represented by θ is $\sigma = e^{\theta}$, this prior choice corresponds to a belief that the measurement noise is around 30%.

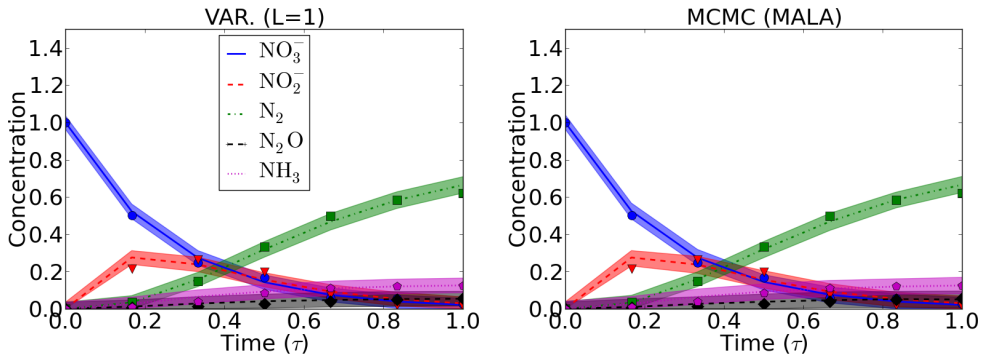

Figure 3: Reaction kinetic model: The '•', '▽', '□', '◇', and '⬠' signs indicate the scaled experimental measurements for NO_3^-, NO_2^-, N_2, N_2O, and NH_3, respectively. The lines and the shaded areas around them correspond to the median and 95% credible intervals of the scaled concentration, u_i, as a function of the scaled time τ. The left plot, shows the results obtained by approximating the posterior of the parameters via the variational approach. The right plot, shows the results obtained via MCMC (MALA).

We solve the problem using the variational approach as outlined in Algorithm 1. To approximate the posterior, we only use one Gaussian, $L = 1$ in Eq. (15). We impose no bounds on $\boldsymbol{\mu}_1 = \boldsymbol{\mu} \in \mathbb{R}^d$. However, we require that the diagonal elements σ_i^2 of the covariance matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma} = \text{diag}\,(\sigma_1, \ldots, \sigma_d)$ are bounded below by 10^{-6} and above by 10^2.

Table 2 compares the variational estimates of the scaled kinetic rate constants, ξ (see Eq. (31)), and the logarithm of the likelihood noise, θ (see Eq. (2)), to the MCMC (MALA) estimates. We see that the mean of the two estimates are in close agreement, albeit the variational approach slightly underestimates the uncertainty of its prediction. However, notice that if we order the parameters in terms of increasing uncertainty, both methods yield the same ordering. Therefore, even though the numerical estimates of the uncertainty differ, the relative estimates of the uncertainty are qualitatively the same. It is worth mentioning at this point, that the variational approach uses only 37 evaluations of the forward model. This is to be contrasted with the thousands of evaluations required so that the MCMC (MALA) estimates converge.

The variational approach with $L = 1$ approximates the posterior of $\boldsymbol{\xi}$ and θ with one multivariate Gaussian distribution with a diagonal covariance. Therefore, the distribution of each one of the components is a Gaussian. Using Eqs. (Eq. (31)) and (Eq. (29)), it is easy to show that the kinetic rate constant k_i follows a log-normal distribution with log-scale parameter $\mu_i - \log(180)$ and shape parameter σ_i. Similarly, we can show that the noise $\sigma = e^{\theta}$ follows a log-normal distribution with local-scale parameter μ_d and shape parameter σ_d. Using the percentiles of these lognormal distributions, we compute the median and the 95% credible intervals of the kinetic rate constants k_i and the noise σ. The results are shown in the third and fourth columns of Table 3. They are in good agreement with the results found in [19] using a MCMC strategy (reproduced in the second column of the same Table 3). An element of our analysis not found

in [19] is the estimation of the measurement noise. Since σ measures the noise of the scaled version of the data $\mathbf{y}$, we see (last line of Table 3) that the measurement noise is estimated to be around 2.15%.

In Fig. 2, we compare the variational posterior (VAR ($L = 1$)) of the scaled kinetic rate constants, κ_i, as well as of the noise of the likelihood, σ, to histograms of the same quantities obtained via MCMC (MALA). Once again, we confirm the excellent agreement between the two methods. In the same figure, we also plot the prior probability density we assigned to each parameter. The prior probability of σ is practically invisible, because it picks at about $\sigma = 0.30$. Given the big disparity between prior and posterior distributions, we see that the result is relatively insensitive to the priors we assign. If, in addition, we take into account that the measurement noise is estimated to be quite small, we conclude that Eq. (26) does a very good job of explaining the experimental observations.

Fig. 3 shows the uncertainty in the scaled concentrations, u_i, as a function of scaled time, τ, obtained by approximating the posterior of the parameters and compares them with the variational approach (left), to the MCMC (MALA) estimate. Again, we notice that the two plots are in very good agreement, albeit the variational approach seems to slightly underestimate the uncertainty.

3.2 Contaminant source identification

We now apply our methodology to a synthetic example of contaminant source identification. We are assuming that we have experimental measurements of contaminant concentrations at specific locations, and we are interested in estimating the location of the contamination source.

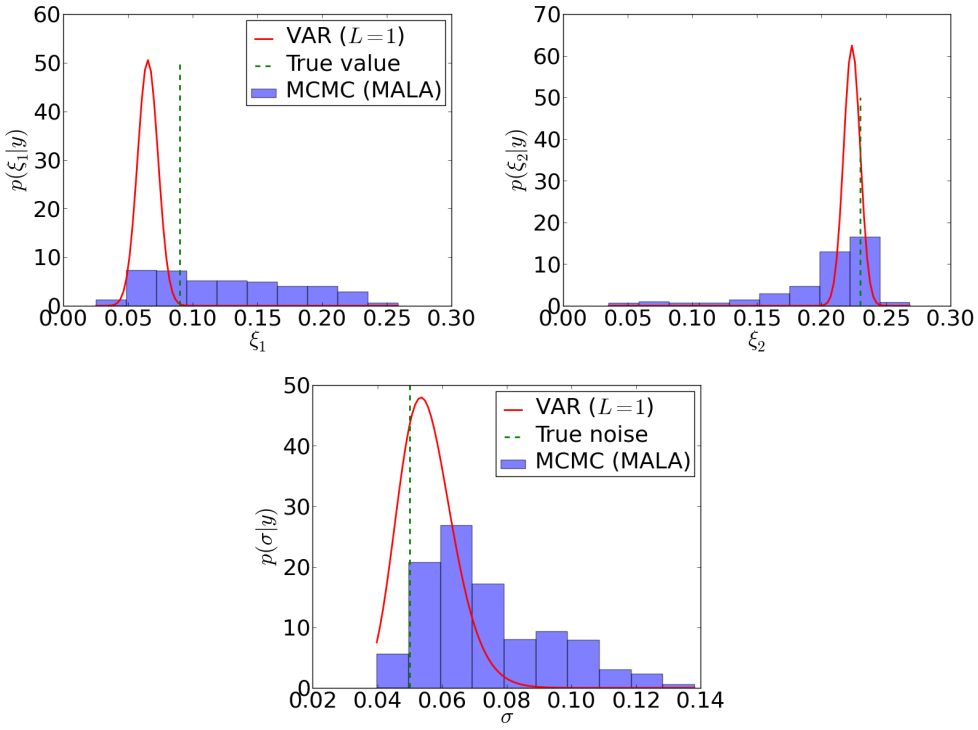

Figure 4: Contaminant source identification, first case: Comparison of the variational posterior (VAR ($L = 1$)) of the source location $\boldsymbol{\xi} = (\xi_1, \xi_2)$ as well as of the likelihood noise σ to the MCMC (MALA) histograms of the same quantities. The true value of each quantity is marked by a vertical, green, dashed line.

The concentration of the contaminant obeys the two-dimensional transport model described by a diffusion equation

$$\frac{\partial u}{\partial t} = \nabla^2 u + g(t, \mathbf{x}, \boldsymbol{\xi}), \ \mathbf{x} \in B, \tag{36}$$

where $B = [0,1]^2$ is the spatial domain and $g(t, \mathbf{x}, \boldsymbol{\xi})$ is the source term. The source term is assumed to have a Gaussian:

$$g(t, \mathbf{s}; \mathbf{x}_s) = g_0 e^{-\frac{|\mathbf{x}-\boldsymbol{\xi}|^2}{2\rho^2}} 1_{[0,T_s]}(t), \tag{37}$$

where $g_0 = \frac{1}{\pi\rho}$ is the strength of the contamination, $\rho = 0.05$ is its spreadwidth, $T_s = 0.3$ is the shutoff time parameter, and $\boldsymbol{\xi}$ is the source center. Therefore, $\boldsymbol{\xi}$ is the only parameter that needs to be identified experimentally. We impose homogeneous Neumann boundary conditions

$$\nabla \mathbf{u} \cdot \mathbf{n} = 0, \ \mathbf{x} \in \partial B \tag{38}$$

and zero initial condition

$$\mathbf{u}(0, \mathbf{x}, \boldsymbol{\xi}) = 0. \tag{39}$$

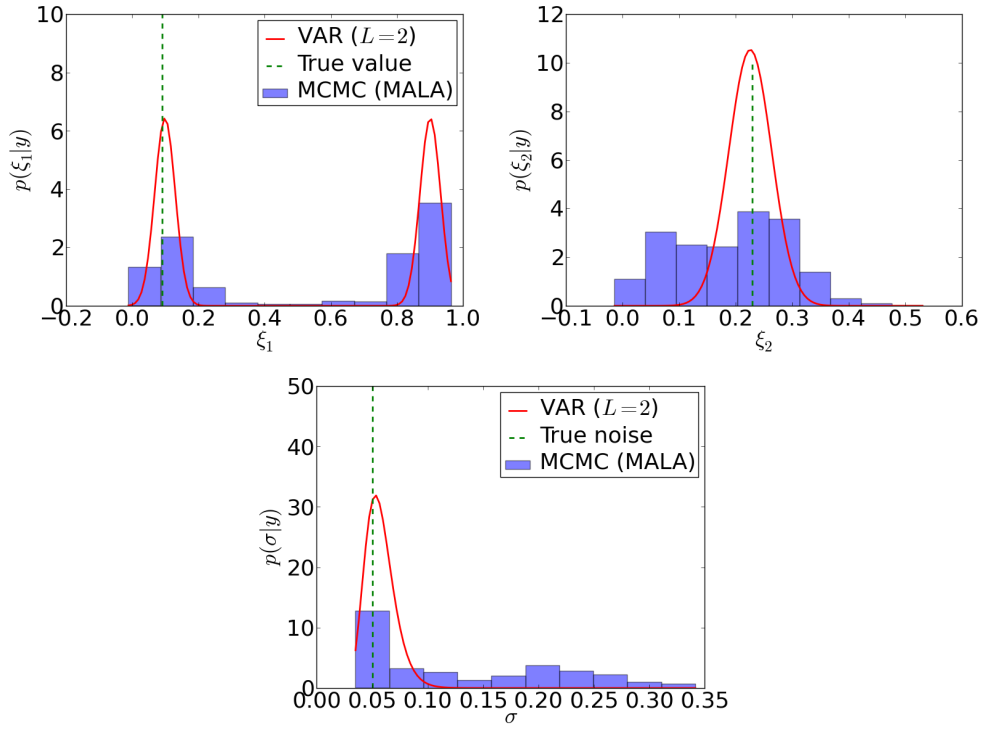

Figure 5: Contaminant source identification, second case: Comparison of the variational posterior (VAR ($L = 1$)) of the source location $\boldsymbol{\xi} = (\xi_1, \xi_2)$ as well as of the likelihood noise σ to the MCMC (MALA) histograms of the same quantities. The true value of each quantity is marked by a vertical, green, dashed line.

We consider two different scenarios. In the first one, measurements of $u(t, \mathbf{x}, \boldsymbol{\xi})$ take place on the four corners of B. On the second one, measurements of $u(t, \mathbf{x}, \boldsymbol{\xi})$ take place on the middle points of the upper and lower boundaries of B. The former results in a unimodal posterior

for $\boldsymbol{\xi}$ and can be approximated with just one Gaussian. The latter results in a bimodal posterior for $\boldsymbol{\xi}$ and requires a mixture of two Gaussians. Eq. (36) is solved via a finite volume scheme implemented using the Python package Fipy[13]. The required gradients of the solution $u(t, \mathbf{x}, \boldsymbol{\xi})$ are obtained by solving a series of PDE's similar to Eq. (36) but with different source terms (see A.6). In both scenarios, we generate synthetic observations, $\mathbf{y}$, by solving Eq. (36) on a 110×110 grid, source $\boldsymbol{\xi}^* = (0.09, 0.23)$, and adding Gaussian noise with standard deviation $\sigma^* = 0.05$. For the forward evaluations needed during the calibration process we use a 25×25 grid and we denote the corresponding solution by $\tilde{\mathbf{u}}(t, \mathbf{s}; \boldsymbol{\xi})$. The prior we use for $\boldsymbol{\xi}$ is uniform,

$$p(\boldsymbol{\xi}) \propto 1_B(\boldsymbol{\xi}), \tag{40}$$

the likelihood is given by Eq. (2), and the log-noise parameter θ of the likelihood has the same prior as the previous example (see Eq. (35)).

First case: Observations at the four corners The synthetic data, $\mathbf{y} \in \mathbb{R}^{16}$ (4 sensors $\times$ 4 measurements) are generated by sampling the 110×110 solution, $u(t, \mathbf{x}) := u(t, \mathbf{x}, \boldsymbol{\xi}^*)$ at the four corners of B, and by adding Gaussian noise with $\sigma^* = 0.05$:

$$\begin{aligned} \mathbf{y} \quad = \quad & \begin{pmatrix} u(t_1, (0,0)), u(t_1, (1,0)), u(t_1, (0,1)), u(t_1, (1,1)), \\ \vdots \\ u(t_4, (0,0)), u(t_4, (1,0)), u(t_4, (0,1)), u(t_4, (1,1)) \end{pmatrix} \\ & + \text{noise}. \end{aligned} \tag{41}$$

The corresponding forward model generated by the 25×25 solution, $\tilde{u}(t, \mathbf{x}, \boldsymbol{\xi})$, is given by:

$$\mathbf{f}(\boldsymbol{\xi}) = \begin{pmatrix} \tilde{u}(t_1, (0,0), \boldsymbol{\xi}), \tilde{u}(t_1, (1,0), \boldsymbol{\xi}), \tilde{u}(t_1, (0,1), \boldsymbol{\xi}), \tilde{u}(t_1, (1,1), \boldsymbol{\xi}), \\ \vdots \\ \tilde{u}(t_4, (0,0), \boldsymbol{\xi}), \tilde{\mathbf{u}}(t_4, (1,0), \boldsymbol{\xi}), \tilde{u}(t_4, (0,1); \boldsymbol{\xi}), \tilde{u}(t_4, (1,1), \boldsymbol{\xi}) \end{pmatrix}. \tag{42}$$

Fig. 4 compares the posteriors obtained via the variational approach with $L = 1$ Gaussian in Eq. (15) to those obtained via MCMC (MALA). The true value of each parameter is indicated by a vertical, green, dashed line. It is worth noting at this point, that the variational approach required 48 forward model evaluations as opposed to thousands required by MCMC (MALA). In real time, the variational approach took about 15 minutes on a single computational node, while the MCMC (MALA) required 3 days on the same node. Notice that the posterior cannot identify the true source exactly. This is due to the 5% noise that we have added in the synthetic data. The result of such noise is always to broaden the posterior. We see that the variational approach does a good job in identifying an approximate location for the source $\boldsymbol{\xi}$ as well as estimating the noise level σ. However, we notice once again that it underestimates the true uncertainty.

Second case: Observations at the middle point of the upper and lower boundaries The synthetic data, $\mathbf{y} \in \mathbb{R}^8$ (2 sensors $\times$ 4 measurements) are generated by sampling the 110×110 solution, $u(t, \mathbf{x}) := u(t, \mathbf{x}, \boldsymbol{\xi}^*)$ at the upper and lower boundaries of B, and by adding Gaussian noise with $\sigma^* = 0.05$:

$$\begin{aligned} \mathbf{y} \quad = \quad & \begin{pmatrix} u(t_1, (0.5,0)), u(t_1, (0.5,1)), \\ \vdots \\ u(t_4, (0.5,0)), u(t_4, (0.5,1)) \end{pmatrix} \\ & + \text{noise}. \end{aligned} \tag{43}$$

The corresponding forward model generated by the 25×25 solution, $\tilde{u}(t, \mathbf{x}, \boldsymbol{\xi})$, is given by:

$$\mathbf{f}(\boldsymbol{\xi}) = \begin{pmatrix} \tilde{u}(t_1, (0.5, 0), \boldsymbol{\xi}), \tilde{u}(t_1, (0.5, 1), \boldsymbol{\xi}), \\ \vdots \\ \tilde{u}(t_4, (0.5, 0), \boldsymbol{\xi}), \tilde{\mathbf{u}}(t_4, (0.5, 1), \boldsymbol{\xi}) \end{pmatrix}. \tag{44}$$

Fig. 5 compares the posteriors obtained via the variational approach with $L = 2$ Gaussians in Eq. (15) to those obtained via MCMC (MALA). The true value of each parameter is indicated by a vertical, green, dashed line. It is worth noting at this point, that the variational approach required only 62 forward model evaluations. Using symmetry arguments, it is easy to show that data generated by solving Eq. (36) with a source located at (ξ_1, ξ_2) look identical to the data that can be generated from a source located at $(1 - \xi_1, \xi_2)$. As a result, the posterior distribution is bimodal. Therefore, we expect common MCMC methodologies to have a hard time dealing with this problem. The reason is that once the MCMC chain visits one of the modes, it is very unlikely that it will ever leave it to visit the other mode. In reality, it is impossible to visit the other mode unless a direct jump is proposed. The reason our MCMC (MALA) scheme works is because we have handpicked a proposal step that does allow for occasional jumps from one mode to the other. On the other hand, we see that the variational approach with $L = 2$ Gaussians in Eq. (15) can easily deal with bimodal (or multimodal) posteriors. However, there are a few details that need to be mentioned here. Firstly, one needs to use an L greater than or equal to the true number of modes of the posterior. Since, the latter is unknown, a little bit of experimentation would be required in a real problem. Secondly, even if the true L is used, Algorithm 1 might still find fewer modes than the true number. For example, in our numerical experiments we have noticed that if $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of Eq. (15) with $L = 2$ are initially very close together, they are both attracted by the same mode. We believe that this is an artifact introduced by the Taylor approximation to the joint probability function $\mathcal{L}[q]$ (see Eq. (22) and Eq. (23)), and, in particular, of its local nature. In our example, a few random initializations of the $\boldsymbol{\mu}_i$'s are enough to guarantee the identification of both posterior modes.

4 CONCLUSIONS

We presented a novel approach to inverse problems that provides an optimization perspective to the Bayesian point of view. In particular, we used information theoretic arguments to derive an optimization problem that targets the estimation of the posterior within the the class of mixtures of Gaussians. The scheme proceeds by postulating that the "best" approximate posterior is the one that exhibits the minimum information loss (relative entropy) within the class of candidate posteriors. We showed how the minimization of the information loss is equivalent to the maximization of a lower bound to the evidence (normalization constant of the posterior). Since the derived lower bound was a computationally intractable quantity, we derived a crude approximation to it that requires the gradients of the forward model with respect to the input variables that we want to infer.

We demonstrated the efficacy of our method to solve inverse problems with just a few forward model evaluations in two numerical examples: 1) the estimation of the kinetic rate constants in a catalysis system, and 2) the identification of the contamination source in a simple diffusion problem. The performance of the scheme was compared to that of a state of the art MCMC technique (MALA) and was found to be satisfactory, albeit slightly underestimating the uncertainty. The scheme was able to solve both inverse problems with a fraction of the computational cost. In particular, our approach required around 50 forward model evaluations

as opposed to the tens of thousands that are required by MCMC.

The variational approach seems to open up completely new ways of solving stochastic inverse problems. The scope of the approach is much wider than the particular techniques used in this paper. Just as a indication, the following are some of the research directions that we plan to pursue in the near future: 1) Derive alternative -more accurate- approximations to the lower bound of the evidence; 2) Experiment with dimensionality reduction ideas that would allow us to carry out the variational optimization in high-dimensional problems; 3) Derive stochastic algorithms for maximizing the lower bound without invoking any approximations. We believe that the variational approach has the potential of making stochastic inverse problems solvable with only a moderate increase in the computational cost as compared to classical approaches.

ACKNOWLEDGEMENTS

I.K. acknowledges financial support through a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (Award IOF-327650). N.Z, as 'Royal Society Wolfson Research Merit Award' holder acknowledges support from the Royal Society and the Wolfson Foundation. N.Z. also acknowledges strategic grant support from EPSRC to the University of Warwick for establishing the Warwick Centre for Predictive Modeling (grant EP/L027682/1). In addition, N.Z. as Hans Fisher Senior Fellow acknowledges support of the Technische Universität München - Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement no. 291763.

A COMPUTATION OF $\mathcal{F}_a[q]$ AND ITS GRADIENT

Algorithm 1 requires the evaluation of the gradient of the approximate ELBO of Eq. (24) with respect to all the parameters of the Gaussian mixture $q(\boldsymbol{\omega})$ of Eq. (15). That is, we must be able to evaluate

$$\mathcal{F}_a[q] = \mathcal{H}_0[q] + \mathcal{L}_a[q], \tag{45}$$

$$\frac{\partial}{\partial\beta}\mathcal{F}_a[q] = \frac{\partial}{\partial\beta}\mathcal{H}_0[q] + \frac{\partial}{\partial\beta}\mathcal{L}_a[q], \tag{46}$$

for $\beta = w_i, \mu_{ij} = (\boldsymbol{\mu}_i)_j, \Sigma_{ijk} = (\boldsymbol{\Sigma}_i)_{jk}$ for $i = 1, \ldots, L$, and $j, k = 1, \ldots, d$ and $a = 0, 2$.

A.1 Computation of $\mathcal{H}_0[q]$ and its gradient

The computations relative to $\mathcal{H}_0[q]$ are:

$$\mathcal{H}_0[q] = -\sum_{i=1}^{L} w_i \log(q_i), \tag{47}$$

$$\frac{\partial}{\partial w_i}\mathcal{H}_0[q] = -\log(q_i) - \sum_{r=1}^{L} \frac{w_r N_{ri}}{q_r}, \tag{48}$$

$$\frac{\partial}{\partial \mu_{ij}}\mathcal{H}_0[q] = -w_i \sum_{r=1}^{L} w_r N_{ri} A_{rij} \left(\frac{1}{q_i} + \frac{1}{q_r}\right), \tag{49}$$

$$\frac{\partial}{\partial \Sigma_{ijk}} H_0[q] = \frac{1}{2} w_i \sum_{r=1}^{L} w_r N_{ri} B_{rijk} \left(\frac{1}{q_i} + \frac{1}{q_r}\right), \tag{50}$$

where, in order to simplify the notation, we have used the following intermediate quantities:

$$N_{ri} = \mathcal{N}\left(\boldsymbol{\mu}_r|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_i\right), \tag{51}$$

$$q_i = \sum_{r=1}^{L} w_r N_{ri}, \tag{52}$$

$$A_{rij} = \sum_{s=1}^{d} \left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_i\right)^{-1}_{js} \left(\mu_{rs} - \mu_{is}\right), \tag{53}$$

$$B_{rijk} = \left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_i\right)^{-1}_{jk} + A_{rij}A_{rik}. \tag{54}$$

A.2 Computation of $\mathcal{L}_a[q]$ and its gradients

The computations relative to the $\mathcal{L}_a[q]$ part are:

$$\mathcal{L}_0[q] = \sum_{i=1}^{L} w_i C_i, \tag{55}$$

$$\frac{\partial}{\partial w_i}\mathcal{L}_0[q] = C_i, \tag{56}$$

$$\frac{\partial}{\partial \mu_{ij}}\mathcal{L}_0[q] = D_{ij}, \tag{57}$$

$$\frac{\partial}{\partial \Sigma_{ijk}}\mathcal{L}_0[q] = 0, \tag{58}$$

$$\mathcal{L}_2[q] = \mathcal{L}_0[q] + \frac{1}{2}\sum_{i=1}^{L} w_i \sum_{j,k=1}^{d} \Sigma_{ijk}E_{ijk}, \tag{59}$$

$$\frac{\partial}{\partial w_i}\mathcal{L}_2[q] = \frac{\partial}{\partial w_i}\mathcal{L}_0[q] + \frac{1}{2}\sum_{j,k=1}^{d} \Sigma_{ijk}E_{ijk}, \tag{60}$$

$$\frac{\partial}{\partial \Sigma_{ijk}}\mathcal{L}_2[q] = \frac{\partial}{\partial \Sigma_{ijk}}\mathcal{L}_0[q] + \frac{1}{2}w_i E_{ijk}, \tag{61}$$

where, in order to simplify the notation, we have used the following intermediate quantities:

$$J(\boldsymbol{\omega}) = \log p(\mathbf{y}, \boldsymbol{\omega}) = \log p(\mathbf{y}|\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}), \tag{62}$$

$$C_i = J(\boldsymbol{\mu}_i), \tag{63}$$

$$D_{ij} = \frac{\partial}{\partial \omega_j} J(\boldsymbol{\mu}_i), \tag{64}$$

$$E_{ijk} = \frac{\partial^2}{\partial \omega_j \omega_k} J(\boldsymbol{\mu}_i). \tag{65}$$

We do not provide the derivatives of $\mathcal{L}_2[q]$ with respect to μ_{ijk} because they are not needed in Algorithm 1. The joint probability function $J(\boldsymbol{\omega})$ of Eq. (62) depends on the details of the likelihood, the prior, and the underlying forward model. We discuss the computation of Eq. (62), Eq. (64) , and Eq. (65) in A.3.

A.3 Computing the derivatives of $J(\omega)$

In this section, we show how the gradient of the forward model appears through the differentiation of the joint probability density $J(\boldsymbol{\omega})$ of Eq. (62). Recall (see beginning of Sec. 2) that

$\boldsymbol{\omega} = (\boldsymbol{\xi}, \boldsymbol{\theta})$, where $\boldsymbol{\xi}$ are the parameters of the forward model $\mathbf{f}(\boldsymbol{\xi}) \in \mathbb{R}^m$, and $\boldsymbol{\theta}$ the parameters of the likelihood function of Eq. (1). Therefore, let us write:

$$
\begin{aligned}
J(\boldsymbol{\omega}) = J(\boldsymbol{\xi}, \boldsymbol{\theta}) &= L(\mathbf{y}, \mathbf{f}(\boldsymbol{\xi}), \boldsymbol{\theta}) + P_\xi(\boldsymbol{\xi}) + P_\theta(\boldsymbol{\theta}), && (66)\\
L(\mathbf{y}, \mathbf{f}(\boldsymbol{\xi}), \boldsymbol{\theta}) &= \log p(\mathbf{y}|\mathbf{f}(\boldsymbol{\xi}), \boldsymbol{\theta}), && (67)\\
P_\xi(\boldsymbol{\xi}) &= \log p(\boldsymbol{\xi}), && (68)\\
P_\theta(\boldsymbol{\theta}) &= \log p(\boldsymbol{\theta}). && (69)
\end{aligned}
$$

Using the chain rule, we have:

$$
\begin{aligned}
\frac{\partial J}{\partial \xi_j} &= \sum_{s=1}^{d_y} \frac{\partial L}{\partial f_s}\frac{\partial f_s}{\partial \xi_j} + \frac{\partial P_\xi}{\partial \xi_j}, && (70)\\
\frac{\partial J}{\partial \theta_j} &= \frac{\partial L}{\partial \theta_j} + \frac{\partial P_\theta}{\partial \theta_j}, && (71)\\
\frac{\partial^2 J}{\partial \xi_j \partial \xi_j} &= \sum_{s,r=1}^{d_y} \frac{\partial^2 L}{\partial f_r \partial f_s}\frac{\partial f_r}{\partial \xi_k}\frac{\partial f_s}{\partial \xi_j} + \sum_{s=1}^{d_y} \frac{\partial L}{\partial f_s}\frac{\partial^2 f_s}{\partial \xi_j \partial \xi_k} + \frac{\partial^2 P_\xi}{\partial \xi_j \partial \xi_k}, && (72)\\
\frac{\partial^2 J}{\partial \theta_i \partial \theta_j} &= \frac{\partial^2 L}{\partial \theta_i \partial \theta_j} + \frac{\partial^2 P_\xi}{\partial \theta_i \partial \theta_j}, && (73)\\
\frac{\partial^2 J}{\partial \xi_j \partial \theta_k} &= \sum_{s=1}^{d_y} \frac{\partial^2 L}{\partial \theta_k \partial f_s}\frac{\partial f_s}{\partial \xi_j}. && (74)
\end{aligned}
$$

Therefore, the Jacobian and the Hessian of the forward model are required. However, as is obvious by close inspection of Eqs. (Eq. (59)), (Eq. (60)), and (Eq. (61)), if the covariance matrices of the mixture $q(\boldsymbol{\omega})$ of Eq. (15) are diagonal, then only the diagonal elements of the Hessian of the forward model are essential. This is the approach we follow in our numerical examples.

A.4 Isotropic Gaussian likelihood

In both our numerical examples, we use the isotropic Gaussian likelihood defined in Eq. (2). Its logarithm is:

$$
L(\mathbf{y}, \mathbf{f}(\boldsymbol{\xi}), \theta) = \log \mathcal{N}\left(\mathbf{y}|\mathbf{f}(\boldsymbol{\xi}), e^{2\theta}\mathbf{I}\right). \quad (75)
$$

The required gradients are:

$$
\begin{aligned}
\frac{\partial L}{\partial f_r} &= e^{-2\theta}\left(y_r - f_r(\boldsymbol{\xi})\right),\\
\frac{\partial L}{\partial \theta} &= e^{-\theta}\left(\| \mathbf{y} - \mathbf{f}(\boldsymbol{\xi}) \|_2^2 \, e^{-2\theta} - d_\xi\right),\\
\frac{\partial^2 L}{\partial \theta^2} &= e^{-\theta}\left(d_\xi - 3 \| \mathbf{y} - \mathbf{f}(\boldsymbol{\xi}) \|_2^2 \, e^{-2\theta}\right),\\
\frac{\partial^2 L}{\partial f_r \partial f_s} &= -e^{-2\theta},\\
\frac{\partial^2 L}{\partial \theta \partial f_r} &= -2e^{-3\theta}\left(y_r - f_r(\boldsymbol{\xi})\right).
\end{aligned}
$$

A.5 Derivatives of a dynamical system

Assume that $\mathbf{u}(t, \boldsymbol{\xi}) \in \mathbb{R}^{d_u}$ is the solution of the following initial value problem:

$$\begin{array}{rl} \dot{\mathbf{u}} & = \mathbf{g}(\mathbf{u}, t, \boldsymbol{\xi}), \\ \mathbf{u}(0) & = \mathbf{u}_0(\boldsymbol{\xi}), \end{array} \tag{76}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{d_\xi}$ are parameters. Using the chain rule, one can show that the derivatives of $\mathbf{u}(t, \boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$, $v_{ij} = \frac{\partial u_i}{\partial \xi_j}$, satisfy the following initial value problem:

$$\begin{array}{rl} \dot{v}_{ij} & = \sum_{r=1}^{d_u} \frac{\partial g_i}{\partial u_r} v_{rj} + \frac{\partial g_i}{\partial \xi_j}, \\ v_{ij}(0) & = \frac{\partial u_0}{\partial \xi_j}. \end{array} \tag{77}$$

The second derivatives $w_{ijk} = \frac{\partial^2 u_i}{\partial \xi_j \partial \xi_k}$, satisfy the following initial value problem:

$$\begin{array}{rl} \dot{w}_{ijk} & = \sum_{r=1}^{d_u} \frac{\partial g_i}{\partial u_r} w_{rjk} + \sum_{r,s=1}^{d_u} \frac{\partial^2 g}{\partial u_r \partial u_s} v_{rj} v_{sk} + \frac{\partial^2 g_i}{\partial \xi_j \partial \xi_k}, \\ w_{ijk}(0) & = \frac{\partial^2 u_0}{\partial \xi_j \partial \xi_k}. \end{array} \tag{78}$$

The numerical strategy for solving these systems is quite simple. First, we solve Eq. (76) using an explicit runge-kutta method of order (4)5. Then, we use the solution as forcing in Eq. (77) to find the gradient. Finally, both the solution and the gradient are used as forcing in Eq. (78).

A.6 Derivatives of the diffusion equation

Let $u(t, \mathbf{s}, \boldsymbol{\xi}) \in \mathbb{R}^{d_u}$ be the solution of the partial differential equation Eq. (36). The derivatives $v_i = \frac{\partial u}{\partial \xi_i}$ satisfy the partial differential equation

$$\frac{\partial v_i}{\partial t} = \nabla^2 v_i + \frac{\partial g(t, \mathbf{x}, \boldsymbol{\xi})}{\partial \xi_i}. \tag{79}$$

Similarly the second derivatives $w_{ij} = \frac{\partial^2 u}{\partial \xi_i \partial \xi_j}$ satisfy

$$\frac{\partial w_{ij}}{\partial t} = \nabla^2 w_{ij} + \frac{\partial^2 g(t, \mathbf{x}, \boldsymbol{\xi})}{\partial \xi_i \partial \xi_j}. \tag{80}$$

Equations (36), (79), and (80) are solved numerically using the same space- and time-discretization and the finite volume solver provided by FiPy [13]. The only thing that changes is the source term.

REFERENCES

[1] Yves F. Atchadé. An adaptive version for the metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability*, 8:235–254, 2006.

[2] I. Bilionis and P.S. Koutsourelakis. Free energy computations by minimization of Kullback–Leibler divergence: An efficient adaptive biasing potential method for sparse representations. *Journal of Computational Physics*, 231(9):3849–3870, 2012.

[3] I Bilionis and N Zabaras. Solution of inverse problems with limited forward solver evaluations: a bayesian perspective. *Inverse Problems*, 30(1):015004, 2014.

[4] Ilias Bilionis, Beth A. Drewniak, and Emil M. Constantinescu. Crop physiology calibration in CLM. *Geoscientific Model Development (under review)*, 2014.

[5] Ilias Bilionis and Nicholas Zabaras. A stochastic optimization approach to coarse-graining using a relative-entropy framework. *The Journal of Chemical Physics*, 138(4):–, 2013.

[6] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization, 1995.

[7] Peng Chen, Nicholas Zabaras, and Ilias Bilionis. Uncertainty Propagation using Infinite Mixture of Gaussian Processes and Variational Bayesian Inference, in press. *Journal of Computational Physics*, 2014.

[8] A. Fichtner. *Full Seismic Waveform Modelling and Inversion*. Advances in Geophysical and Environmental Mechanics and Mathematics. Springer, 2010.

[9] Andreas Fichtner. Full Seismic Waveform Modelling and Inversion, 2011.

[10] Charles W. Fox and Stephen J. Roberts. A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38(2):85–95, June 2011.

[11] Samuel Gershman, Matthew D. Hoffman, and David M. Blei. Nonparametric variational inference. *CoRR*, abs/1206.4665, 2012.

[12] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition*. SIAM e-books. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2008.

[13] J.E. Guyer, D. Wheeler, and J.A. Warren. Fipy: Partial differential equations with python. *Computing in Science and Engineering*, 11(3):6–15, 2009.

[14] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[15] Mary C. Hill and Claire R. Tiedeman. *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*. Wiley, 2006.

[16] Marco F. Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck. On entropy approximation for gaussian mixture random vectors. In *Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Seoul, Sdkorea, August 2008.

[17] E.T. Jaynes and G.L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[18] Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 1 edition, December 2002.

[19] I Katsounaros, M Dortsiou, C Polatides, S Preston, T Kypraios, and G Kyriacou. Reaction pathways in the electrochemical reduction of nitrate on tin. *Electrochimica Acta*, 71:270–276, 2012.

[20] M.C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 425–464, 2001.

[21] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[22] Youssef Marzouk and Dongbin Xiu. A Stochastic Collocation Approach to Bayesian Inference in Inverse Problems. *Communications in Computational Physics*, 6(4):826–847, 2009.

[23] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

[24] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[25] Ionel Navon M. Data assimilation for numerical weather prediction: A review. In SeonK. Park and Liang Xu, editors, *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, pages 21–65. Springer Berlin Heidelberg, 2009.

[26] R.-E. Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, November 2006.

[27] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 09 1951.

[28] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2005.