Fast Bayesian Inversion for high dimensional inverse problems

Benoit Kugler, Florence Forbes, Sylvain Douté

Fast Bayesian Inversion for high dimensional inverse problems

Benoit Kugler[1,2,*], **Florence Forbes**[1,†] **and Sylvain Douté**[2]

[1] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France, e-mail:* [*] benoit.kugler@inria.fr; [†] florence.forbes@inria.fr

[2] *Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France, e-mail:* sylvain.doute@univ-grenoble-alpes.fr

Abstract: We investigate the use of learning approaches to handle Bayesian inverse problems in a computationally efficient way when the signals to be inverted are high dimensional and in large number. We propose a tractable inverse regression approach which has the advantage to produce full probability distributions as approximations of the target posterior distributions. In addition to provide confidence indices on the predictions, these distributions allow a better exploration of inverse problems when multiple equivalent solutions exist. We then show how these distributions can be used for further refined predictions using importance sampling, while also providing a way to carry out uncertainty level estimation if necessary. The relevance of the proposed approach is illustrated both on simulated and real data in the context of a physical model inversion in planetary remote sensing.

Keywords and phrases: Inverse problems, High dimension, Bayesian analysis, Mixtures of Gaussians, Importance sampling, Remote sensing, Planetary science.

1. Introduction

A wide class of problems from medical imaging (Frau-Pascual et al., 2014; Mesejo et al., 2016; Lemasson et al., 2016; Nataraj et al., 2018) to astrophysics (Chiancone, Forbes and Girard, 2017; Deleforge et al., 2015; Bernard-Michel et al., 2007; Schmidt and Fernando, 2015) can be formulated as inverse problems (Tarantola, 2005; Giovannelli and Idier, 2015). An inverse problem refers to a situation where one aims at determining the causes of a phenomenon from experimental observations of its effects. The resolution of such a problem generally starts by the so-called *direct or forward* modelling of the phenomenon. It theoretically describes how input parameters $\mathbf{x} \in \mathcal{X}$ are translated into effects $\mathbf{y} \in \mathcal{Y}$. Then from experimental observations of these effects, the goal is to find the parameters values that best explain the observed measures.

Typical situations or constraints that can be encountered in practice are that 1) both direct and inverse relationships are (highly) non-linear, *e.g.* the direct model is available but is a (complex) series of ordinary differential equations as in Mesejo et al. (2016); Hovorka et al. (2004); 2) the observations $\mathbf{y}$ are high-dimensional because they represent signals in time or spectra, as in Schmidt and Fernando (2015); Bernard-Michel et al. (2009); Ma et al. (2013); 3) many such high-dimensional observations are available and the application requires a very large number of inversions, *e.g.* Deleforge et al. (2015); Lemasson et al. (2016); 4) the parameters $\mathbf{x}$ to be predicted is itself multi-dimensional with correlated dimensions so that predicting its components independently is sub-optimal, *e.g.* when there are known constraints such as their sum is one like for concentrations or probabilities (Deleforge et al., 2015; Bernard-Michel et al., 2009).

The most standard resolution approaches are based on optimization techniques, minimizing a cost function between observed and modelled effects possibly augmented with regularization terms

(*e.g.* a priori knowledge). However, such methods cannot generally handle massive inversions of high dimensional data and quantify uncertainty. Moreover, the choice of the regularization parameter is not obvious and requires application dependent fine tuning.

Beyond optimization techniques, we propose to investigate learning and regression approaches which are less commonly used to solve inverse problems. The main principle is to transfer the cost of individual inversions to the estimation of an inversion operator that once learned provides multiple predictions at low cost. In addition, to account for uncertainty in a principled manner, we consider Bayesian inversion techniques. They provide a full posterior probability distribution as the output of the inversion. However, Bayesian inversion is often impaired by an intractable normalizing constant in the Bayes formula and requires the use of sampling intensive approaches such as Markov Chain Monte Carlo (MCMC) techniques (Robert and Casella, 2004) or Approximate Bayesian Computation (ABC) techniques (Sisson, Fan and Beaumont, 2018). MCMC and ABC procedures provide samples of $\mathbf{x}$ values that follow the posterior distribution. These samples can be used to compute point estimates of the parameters or more generally to get an empirical approximation of the posterior probability distribution function (pdf). Despite coming with theoretical guaranties and being used for example in remote sensing (Schmidt and Fernando, 2015), medical imaging (Bertrand et al., 2001) or geology (Martin et al., 2012), a limitation of such sampling-based approaches is their cost. A large number of samples has to be simulated and this for each $\mathbf{y}$ to be inverted. This is problematic when the number and dimension of $\mathbf{y}$ increase. Although more and more approaches address this issue, *e.g.* Bardenet, Doucet and Holmes (2014); Izbicki, Lee and Pospisil (2019), sampling techniques do not scale easily to high dimensional settings.

In this work, we develop an approach, referred to as *fast Bayesian inversion*. Assuming parameters and effects are random variables $\mathbf{X} \in \mathcal{X}$ and $\mathbf{Y} \in \mathcal{Y}$, we consider a data set $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n), n = 1 : N\}$, built from the forward model via a data generating model or directly from observed realizations of $\mathbf{X}$ and $\mathbf{Y}$ if available. An inversion operator is learned from $\mathcal{D}$ via a parametric probability distribution $p(\mathbf{x} \mid \mathbf{y}; \theta)$. The model $p(\mathbf{x} \mid \mathbf{y}; \hat{\theta})$ that best fits the data is selected in a family of so-called Gaussian Locally Linear Mapping (GLLiM) models (Deleforge, Forbes and Horaud, 2015). Indeed the latter have the ability to capture non linear relationships in a tractable manner based on flexible mixtures of Gaussian distributions. The estimated parameter $\hat{\theta}$ captures the relationship between $\mathbf{X}$ and $\mathbf{Y}$ as a whole and does not depend on a specific observed $\mathbf{y}$ to be inverted. The learned model $p(\mathbf{x} \mid \mathbf{y}; \hat{\theta})$ can be used as an approximation of the true posterior density denoted by $p_0(\mathbf{x} \mid \mathbf{y})$. Its formulation allows to compute moments, modes etc. straightforwardly so that for each new $\mathbf{y}$ to be inverted we can derive good candidates values for $\mathbf{x}$ at low cost. A simple solution is to estimate $\mathbf{x}$ as the posterior mean, that is the mean of $p(\mathbf{x} \mid \mathbf{y}; \hat{\theta})$. However, this may not be a good prediction when there exist several solutions. In this latter case, the modes of $p(\mathbf{x} \mid \mathbf{y}; \hat{\theta})$ are likely to be a better choice. Overall, we propose to choose between candidate solutions using some reconstruction criterion and to replace a costly continuous search in the parameter space by a discrete choice among a small finite number of candidates solutions (*e.g.* the posterior mean and posterior modes).

We show that such a parametric approximation and the consequent choice of the posterior mean of $p(\mathbf{x} \mid \mathbf{y}; \hat{\theta})$ generally provides satisfying predictions. However, they may not be as accurate as predictions generated by non parametric MCMC or ABC methods when they are feasible. We therefore investigate the augmentation of the parametric approximation with some subsequent importance sampling to get an improved approximation while maintaining tractability. We keep the principle of proposing a discrete number of candidate solutions for efficiency but each of them is refined using an additional importance sampling step.

As for the design of the data set $\mathcal{D}$, we present an Expectation Maximization (EM) algorithm to estimate the appropriate level of uncertainty to use in the data generating model when this level is not given by the experts.

The rest of the paper is organized as follows. In Section 2, we briefly review other fast inversion procedures and justify our Bayesian choice. In Section 3, we recall the main ingredients of the GLLiM parametric regression model used to learn an approximation of $p_0(\mathbf{x} \mid \mathbf{y})$. In Section 4, we specify different ways to exploit the GLLiM output for prediction using means or modes, while in section 5, we propose a further improvement using importance sampling in cases where the likelihood of the forward model is available. The EM algorithm for estimating the uncertainty level is given in Section 6. Illustrations are given in Section 7 with the proposed approach tested on synthetic data and on a challenging real inverse problem in planetary remote sensing. A discussion and conclusion end the paper.

2. Alternative fast inversion procedures

Among alternatives to standard optimization that may provide low cost inversions are *grid search* approaches used for example in remote sensing (Darvishzadeh, Matkan and Ahangar, 2012) and medical imaging (Zhao et al., 2016; Lemasson et al., 2016). The optimization step is replaced by a simpler look-up or matching operation. A large data set $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n), n = 1 : N\}$ is generated (stored or computed on the fly) by running the theoretical model for many different parameter values corresponding to a grid in the full $\mathcal{X}$ space. Note that if $\mathcal{D}$ is available from direct observation, the precise knowledge of the theoretical model is not necessary. Inverting the model on an observed $\mathbf{y}$ consists of comparing $\mathbf{y}$ to the $\mathbf{y}_n$'s in $\mathcal{D}$ in order to find the best match (the nearest neighbor) according to a similarity score. The solution is then set to the parameters $\mathbf{x}_n$ associated to this best match. The speed gain is significant in comparison to traditional optimization methods as retrieving a value from memory is often faster than undergoing an expensive computation. The sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is fixed and does not depend on $\mathbf{y}$ but for each new $\mathbf{y}$, the matching scores to all $\mathbf{y}_n$'s have to be computed. It follows that grid search approaches are prone both to solution instability and intractability in high dimensions (Bernard-Michel et al., 2009).

More recently, deep learning methods have been extensively applied to inverse problems (see Arridge et al. (2019) for a survey). The idea is to replace the theoretical inverse function by an approximate one, deduced from training samples. The learning strategy has been proposed by several groups using deep learning tools (Virtue, Yu and Lustig, 2017; Hoppe et al., 2017; Cohen et al., 2018; Balsiger et al., 2018; Barbieri et al., 2018; Song et al., 2019; Golbabaee et al., 2019). Like with grid search approaches, the prediction with neural networks is fast, since the heavy computations of the first step are done offline. Despite recent successes in classification tasks, neural networks still raise questions. First, the number of parameters of this type of learning machines is so high that the former are very difficult to interpret. In the case of multiple solutions, the inverse model is not even a function. Second, it has been shown that neural networks may be highly non robust to noisy observations. The work in Szegedy et al. (2014) presents striking examples. Third this inversion technique does not generally estimate uncertainties on the solution. Finally, Adler and Öktem (2017) points out that the learning step might require a lot of data, especially if the observation space is high dimensional. Deep learning methods overall require more computational resources.

Other learning or regression methods adapted to high dimensions include inverse regression methods, *i.e.* sliced inverse regression (Li, 1991), partial least squares (Cook and Forzani, 2019), approaches based on mixtures of regressions with different variants, *e.g.* mixtures of experts (Nguyen,

Chamroukhi and Forbes, 2019), cluster weighted models (Ingrassia, Minotti and Vittadini, 2012), and kernel methods (Nataraj et al., 2018). Inverse regression methods are flexible in that they reduce the dimension in a way optimal to the subsequent $\mathbf{y}$ to $\mathbf{x}$ mapping estimation that can itself be carried out by any kind of standard regression tool. In that sense, inverse regression methods are said to be non-parametric or semi-parametric. In Nataraj et al. (2018), the authors propose a regression with an appropriate kernel function to learn the non-linear mapping. The procedure has the advantage to be semi-parametric but a serious limitation is that the model components are optimized in each dimension separately.

Thus, the difficulty is that the above methods are either not specifically designed for high dimensional data or are limited to point-wise predictions with no guaranty or indication of the prediction reliability. In this work, we aim at combining learning and Bayesian approaches to make the best use of the known forward model while taking into account the inherent uncertainties related to the model.

3. Fast Bayesian approach to inverse problems

A natural way to account for uncertainties is to adopt a statistical formulation. Our knowledge on the forward model is encoded in the following *Data generating model*.

3.1. Data generating model

In our inverse problem setting, we call parameters the input values $\mathbf{x}$ that generate the observed effects $\mathbf{y}$. This denomination should not be confused with parameters defining statistical models that we denote by $\boldsymbol{\theta}$ in most cases. The parameters and observations are therefore assumed to be random variables $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^L$ and $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^D$ of dimension L and D respectively where D is usually much greater than L. The forward model is described by a likelihood function linking parameters values $\mathbf{x}$ to the probability of observing some effects $\mathbf{y}$ and denoted by $\mathcal{L}_{\mathbf{x}}(\mathbf{y}) = p_0(\mathbf{y} \mid \mathbf{X} = \mathbf{x})$. We will further assume that the relationship between $\mathbf{X}$ and $\mathbf{Y}$ is described by a known function F and that the uncertainties on the theoretical model are independent on the input parameter $\mathbf{X}$. In other words,

$$\mathbf{Y} = F(\mathbf{X}) + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon}$ is a random variable. For instance, $\boldsymbol{\epsilon}$ is assumed to be a centered Gaussian variable with covariance matrix $\boldsymbol{\Sigma}$, so that $\mathcal{L}_{\mathbf{x}}(\mathbf{y}) = \mathcal{N}(\mathbf{y}; F(\mathbf{x}), \boldsymbol{\Sigma})$, where $\mathcal{N}(\,.\,; F(\mathbf{x}), \boldsymbol{\Sigma})$ denotes the Gaussian pdf with mean $F(\mathbf{x})$ and covariance $\boldsymbol{\Sigma}$. Firstly we assume that $\boldsymbol{\Sigma}$ is given but this constraint is then relaxed with a proposal to estimate $\boldsymbol{\Sigma}$ in Section 6.

To complete the model, we consider a prior distribution on the possible parameter values denoted by $p_0(\mathbf{x})$. This probabilistic formulation allows, according to the Bayesian theorem, to transform an a priori probability distribution into a posterior distribution $p_0(\mathbf{x} \mid \mathbf{y}) \propto \mathcal{L}_{\mathbf{x}}(\mathbf{y})\ p_0(\mathbf{x})$ which incorporates the physical model and the actual observations with their uncertainties. However, even with the simplifying Gaussian likelihood function above, in most cases of interest F is non linear and the Bayesian inversion does not provide an explicit solution for the posterior $p_0(\mathbf{x} \mid \mathbf{y})$ due to an intractable normalizing constant in the Bayes formula.

As an alternative to sampling of the posterior (MCMC or ABC) which does not scale well with dimension and number of inversions to be carried out, we propose to use the data generating model

$(p_0(\mathbf{x}), \mathcal{L}_{\mathbf{x}}(\mathbf{y}))$ for the learning phase of the inversion. For reasonable choices of the two model ingredients, the prior and the likelihood, generating a data set $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n), n = 1 : N\}$ is a straightforward task. Without further knowledge, a common setting is to use a uniform prior on the parameter space and a Gaussian likelihood. From this data set, we can then adopt a regression or learning approach to learn a mapping from the parameter space to the observation space. The main principle is to transfer the cost of individual heavy simulation-based inversions to the learning of a global inverse operator which can then be applied at very little cost to a large number of $\mathbf{y}$ values. To perform this task, we propose to use the Gaussian Locally Linear Mapping (GLLiM) approach described in the next section.

3.2. Parametric posterior approximation with Gaussian mixtures

In the same vein as inverse regression approaches, and in contrast to deep learning approaches mentioned in Section 2, we propose to use the Gaussian Locally Linear Mapping (GLLiM) model (Deleforge, Forbes and Horaud, 2015) that provides a probability distribution selected in a family of mixture of Gaussian distributions $\{p(\mathbf{x}\,|\,\mathbf{y};\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where the mixture parameters are denoted by $\boldsymbol{\theta}$. There have been several extensions and uses of GLLiM, including more robust (Perthame, Forbes and Deleforge, 2018; Tu et al., 2019) and deep (Lathuiliere et al., 2017) versions. However in all these contexts, the focus is on using the model for predictions without fully exploiting the posterior distributions provided by GLLiM.

An attractive approach for modeling non linear data is to use a mixture of linear models. We assume that each $\mathbf{Y}$ is the noisy image of $\mathbf{X}$ obtained from a K-component mixture of affine transformations. This is modeled by introducing a latent variable $Z \in \{1, \ldots, K\}$ such that

$$\mathbf{Y} \quad = \quad \sum_{k=1}^{K} \mathbb{I}_{\{Z=k\}}(\boldsymbol{A}_k\mathbf{X} + \boldsymbol{b}_k + \boldsymbol{\epsilon}_k) \tag{2}$$

where $\mathbb{I}$ is the indicator function, $\boldsymbol{A}_k$ a $D \times L$ matrix and $\boldsymbol{b}_k$ a vector of $\mathbb{R}^D$ that define an affine transformation. Variable $\boldsymbol{\epsilon}_k$ corresponds to an error term which is assumed to be zero-mean and not correlated with $\mathbf{X}$ capturing both the observation noise and the reconstruction error due to the affine approximation. To make the affine transformations local, the latent variable Z should also depend on $\mathbf{X}$.

For the posterior distribution $p(\mathbf{x} \mid \mathbf{y})$ to be easily derived from the likelihood $p(\mathbf{y} \mid \mathbf{x})$, it is important to control the nature of the joint $p(\mathbf{y}, \mathbf{x})$. Once a family of tractable joint distributions is chosen, we can look for one that is compatible with (2). In Deleforge, Forbes and Horaud (2015) the GLLiM model is derived assuming that the joint distribution is a mixture of Gaussian distributions. Using a subscript G to specify the model, it is assumed that $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$ and that $\mathbf{X}$ is distributed as a mixture of K Gaussian distributions specified by $p_G(\mathbf{x}|Z = k) = \mathcal{N}(\mathbf{x}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$, and $p_G(Z = k) = \pi_k$. It follows that the model parameters are $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1:K}$.

One interesting property of such a parametric model is that the mixture setting provides some guaranties that when choosing K large enough it is possible to approximate any reasonable relationship (Nguyen, Chamroukhi and Forbes, 2019) and that both conditional distributions are available

in closed form :

$$p_G(\mathbf{y}|\mathbf{X}=\mathbf{x},\boldsymbol{\theta}) = \sum_{k=1}^{K} \eta_k(\mathbf{x})\mathcal{N}(\mathbf{y};\boldsymbol{A}_k\mathbf{x}+\boldsymbol{b}_k,\boldsymbol{\Sigma}_k) \quad \text{with } \eta_k(\mathbf{x}) = \frac{\pi_k\mathcal{N}(\mathbf{x};\boldsymbol{c}_k,\boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x};\boldsymbol{c}_j,\boldsymbol{\Gamma}_j)} \tag{3}$$

$$p_G(\mathbf{x}|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}^*) = \sum_{k=1}^{K} \eta_k^*(\mathbf{y})\mathcal{N}(\mathbf{x};\boldsymbol{A}_k^*\mathbf{y}+\boldsymbol{b}_k^*,\boldsymbol{\Sigma}_k^*) \quad \text{with } \eta_k^*(\mathbf{y}) = \frac{\pi_k\mathcal{N}(\mathbf{y};\boldsymbol{c}_k^*,\boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K}\pi_j^*\mathcal{N}(\mathbf{y};\boldsymbol{c}_j^*,\boldsymbol{\Gamma}_j^*)}. \tag{4}$$

In (4), a new parametrization $\boldsymbol{\theta}^* = \{\pi{+}k, \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*, \boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*\}_{k=1:K}$ is used to illustrate the similarity between the two conditional distributions (3) and (4). The parameter $\boldsymbol{\theta}^*$ is easily deduced from $\boldsymbol{\theta}$ as follows:

$$\begin{aligned}
\boldsymbol{c}_k^* =& \boldsymbol{A}_k\boldsymbol{c}_k + \boldsymbol{b}_k, \quad \boldsymbol{\Gamma}_k^* = \boldsymbol{\Sigma}_k + \boldsymbol{A}_k\boldsymbol{\Gamma}_k\boldsymbol{A}_k^\top \\
\boldsymbol{\Sigma}_k^* =& \left(\boldsymbol{\Gamma}_k^{-1} + \boldsymbol{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{A}_k\right)^{-1} \\
\boldsymbol{A}_k^* =& \boldsymbol{\Sigma}_k^*\boldsymbol{A}_k^\top\boldsymbol{\Sigma}_k^{-1}, \quad \boldsymbol{b}_k^* = \boldsymbol{\Sigma}_k^*\left(\boldsymbol{\Gamma}_k^{-1}\boldsymbol{c}_k - \boldsymbol{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{b}_k\right) .
\end{aligned} \tag{5}$$

In practice when D is much larger than L, it is more efficient to estimate $\boldsymbol{\theta}$ from the available data $\mathcal{D}$ to then deduce $\boldsymbol{\theta}^*$ and subsequently the conditional distribution of interest (4). The size of $\boldsymbol{\theta}$ can be significantly reduced by choosing constraints on matrices $\boldsymbol{\Sigma}_k$ without oversimplifying the target conditional (4). Typically, diagonal covariance matrices can be used with a drastic gain. To estimate $\boldsymbol{\theta}$ a standard Expectation-Maximization (EM) algorithm can be used. All details are provided in Deleforge, Forbes and Horaud (2015).

Fitting a GLLiM model to $\mathcal{D}$ results in an analytical expression denoted by $p_G(\mathbf{x}|\mathbf{y},\hat{\boldsymbol{\theta}}^*)$ of the form (4) which is a mixture of Gaussian distributions and can be seen as a parametric mapping from $\mathbf{y}$ values to the pdfs on $\mathbf{x}$. The parameter $\hat{\boldsymbol{\theta}}^*$ is the same for all conditional distributions and does not need to be re-estimated for each new $\mathbf{y}$ to be inverted. This is in contrast to sampling procedures like MCMC which would require a new set of samples for each inversion.

A recent result (Nguyen, Chamroukhi and Forbes, 2019), on the density of multiple output mixtures of expert models, justifies a somewhat arbitrary choice of K as soon as it is large enough. Intuitively, highly non-linear F may require a greater K. Automatic model selection procedures can also be used to select K (see Deleforge, Forbes and Horaud (2015)). Alternatively, the choice of K can be guided by the quality of the learned direct model, which only requires a learning data set to be evaluated.

4. Fast inversions and predictions

In applications, the main question is then how to use the approximate posterior $p_G(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}^*)$ to provide predictions or more general information on the value of $\mathbf{x}$ for a given $\mathbf{y}$. We consider mainly prediction by the means and by the modes and investigate below different ways to approximate these quantities.

4.1. Different prediction schemes

4.1.1. Prediction using the posterior mean

The approximate posterior mean $\mathbb{E}_G\left[\mathbf{X} \,|\, \mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^*\right]$ minimizes the mean square error and is given by the mean of (4). For a given $\mathbf{y}$, a first estimator for $\mathbf{x}$ is then

$$\bar{\mathbf{x}}_G(\mathbf{y}) = \mathbb{E}_G\left[\mathbf{X} \,|\, \mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^*\right] = \sum_{k=1}^{K} \eta_k^*(\mathbf{y}) \left(\boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*\right).$$

Its computation is straightforward once the GLLiM model has been learned. As shown later in our experiments and various papers (Deleforge, Forbes and Horaud, 2015; Deleforge et al., 2015; Perthame, Forbes and Deleforge, 2018; Tu et al., 2019), it performs well in several cases. Similarly, other moments can be easily computed. The variance of distribution (4), which accounts for the dispersion around the prediction by the mean, admits the following expression:

$$var_G(\mathbf{y}) = \sum_{k=1}^{K} \eta_k^*(\mathbf{y}) \left[\boldsymbol{\Sigma}_k^* + (\boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*)(\boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*)^T\right] - \bar{\mathbf{x}}_G(\mathbf{y})\, \bar{\mathbf{x}}_G(\mathbf{y})^T.$$

4.1.2. Prediction with the posterior modes

Prediction by the mean is likely to yield good results when the solution is unique and the posterior unimodal. However, it is only a limited summary of the posterior whose full shape may be more informative. When the inverse problem has multiple solutions, the true posterior distribution may reflect that by exhibiting several modes. It follows that the GLLiM approximation may also be multimodal and the prediction by the mean, although a good approximation of the true $E[\mathbf{X}|\mathbf{Y} = \mathbf{y}]$, may provide a low probability solution that is not satisfying for the user. Another problem arises when F has a low sensitivity to part of its variables. Even if the posterior distribution is unimodal, its mean may differ quite largely from its mode.

An alternative then is to look at the modes of the mixture (4) provided by GLLiM. Unfortunately, finding the modes of a Gaussian mixture is not an easy task (Ray and Lindsay, 2005). Heuristics have been proposed starting from the mixture centroids (Carreira-Perpinan, 2000) but in contrast to the mean, there is no analytical formula for the modes and they need to be determined for each new $\mathbf{y}$. The issue of multiple solutions and their localization is highly dependent on $\mathbf{y}$. In practice, their number is not known. The mixture centroids which correspond to the means of the mixture components are not usually good candidates for the modes. However, as suggested in Carreira-Perpinan (2000), they can be good starting points for a fixed-point iterative scheme. When there are a lot of components in the mixture, the density could be explored starting from the centroids with the highest weights to locate the main possible modes using this heuristic scheme. We refer to Carreira-Perpinan (2000) for details. We will not further discuss this issue.

In what follows we propose to address the case of two solutions for $\mathbf{x}$ but the same approach generalizes for more. To account for all mixture components, we consider the following procedure. For any $\mathbf{y}$ of interest, the GLLiM approximation of the posterior $p_G(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta}^*)$ will likely exhibit a majority of low probability components with only a few of them useful for inverting $\mathbf{y}$. Recall that the weight of component k is given by $\eta_k^*(\mathbf{y}) = \eta^*(\mathbf{y})^{-1} \pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)$, $\eta^*(\mathbf{y})$ being a normalisation factor (independent of k). Thus, only components with centroids close to $\mathbf{y}$ and high π_k are likely to

matter. For a simplified, exploitable mixture, the initial K-component Gaussian mixture is modified into a 2-component one. Among the many merging algorithms (see Hennig (2010) for an overview), the algorithm described in Runnalls (2007) is used. This algorithm relies on two steps :

- given two Gaussian components, merge them into a single Gaussian with same weight, mean and variance as the original sum;
- at each iteration, select the two components to be merged by minimizing the Kullback-Leibler divergence between the current mixture and the one that would be obtained by merging these two components.

It follows a final mixture of 2-components with respective weight, mean and covariance denoted by $(\pi_1^y, \boldsymbol{\mu}_1^y, \boldsymbol{\Sigma}_1^y)$ and $(\pi_2^y, \boldsymbol{\mu}_2^y, \boldsymbol{\Sigma}_2^y)$.

The new centroids are considered as prediction candidates and denoted by $\hat{\mathbf{x}}_{centroid,1}(\mathbf{y}) = \boldsymbol{\mu}_1^y$ and $\hat{\mathbf{x}}_{centroid,2}(\mathbf{y}) = \boldsymbol{\mu}_2^y$. It is expected that these centroids are close to the modes of the posterior, but there is no guaranty in the general case. In addition, the merging cost is cubic in K, which might be problematic to maintain computational efficiency. However, this step may be strongly speed up by discarding the none significant Gaussian components. As already pointed out, the posterior is likely to be composed of a majority of very low weighted components, which are the components far away from the observed $\mathbf{y}$. In this case, a preliminary step can be added to merge for instance the 50% less significant components with an expected 8 times speed gain. Alternatively, components can be removed when their weight is below a fixed threshold. A threshold as small as $10e^{-10}$ can already be efficient for this purpose. One important property of the merging algorithm is that the mean and variance of the initial GLLiM mixture are preserved, that is $\bar{\mathbf{x}}_G(\mathbf{y}) = \pi_1^y\ \hat{\mathbf{x}}_{centroid,1}(\mathbf{y}) + \pi_2^y\ \hat{\mathbf{x}}_{centroid,2}(\mathbf{y})$. Hence, if the mean is close to one of the two final centroids, this means that one of the weight π_i^y is close to one while the other one is close to zero, suggesting a unimodal distribution. In contrast, a mean far from both centroids suggests multimodality.

4.1.3. Comparison and selection of predictions

The analysis of the approximate posterior potentially provides a number of possible predictions for $\mathbf{x}$. When F is known, a simple criterion to compare them is to compute the residuals, that is the distance of $F(\mathbf{x})$ to the observed $\mathbf{y}$. The residuals can be compared for each of the three predictions $\bar{\mathbf{x}}_G$, $\hat{\mathbf{x}}_{centroid,1}$ and $\hat{\mathbf{x}}_{centroid,2}$ using the formula, $R(\mathbf{x}) = ||\mathbf{y} - F(\mathbf{x})||_2/||\mathbf{y}||_2$. In a multi-solution scenario, the two centroids should have low residual errors while the mean should have a higher one. Reversely, a high residual error for one of the centroids, combined with a low weight, indicates that it should be discarded. The residual error criterion can be used to assess other values of $\mathbf{x}$.

4.2. Accounting for measurement errors

Independently of the chosen prediction scheme, the use of a unique $\boldsymbol{\theta}^*$ parameter for all inversions provides a great gain when massive inversions are required but it also assumes that the same model is valid for all observations to be inverted and that the dictionary $\mathcal{D}$ is a good representation of them. Although this is a standard assumption, we can point out another useful feature of GLLiM which is to efficiently adapt to known measurement errors. This corresponds to the case where the observed $\mathbf{y}_{obs}$ comes with some covariance matrix $\boldsymbol{\Sigma}_{obs}$ indicating that the data provider was able to measure uncertainty on the observed data. We interpret this additional information as a

measurement error to be distinguished from $\mathbf{\Sigma}$ in (1) which rather accounts for uncertainty on the model. It follows that the observation to be inverted should rather be written as $\mathbf{y} = \mathbf{y}_{obs} + \boldsymbol{\epsilon}_{obs}$ where $\boldsymbol{\epsilon}_{obs}$ is following a centered Gaussian distribution with covariance $\mathbf{\Sigma}_{obs}$. This means that the initial dictionary $\mathcal{D}$ is not fully adapted to invert $\mathbf{y}$ if the data generating process does not account for this additional measurement error. Therefore, another training set $\mathcal{D}_{obs}$ should be simulated and used instead, with a corrected likelihood corresponding to $\mathbf{Y} = F(\mathbf{X}) + \boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{obs}$. Obviously this may be too time consuming if this has to be done for each inversion. Fortunately, it is straightforward to check that the structure of the Gaussian mixture approximation avoid the re-learning of the GLLiM model. Indeed, it suffices to change the estimated $\mathbf{\Sigma}_k$'s into $\mathbf{\Sigma}_k + \mathbf{\Sigma}_{obs}$ and to report this change when computing $\boldsymbol{\theta}^*$ in (5) and the corresponding $p_G(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^*)$. It should be noted that the uncertainty on the model and the measurement errors combine in our inversion scheme. This is actually a generic characteristic of the Bayesian formalism applied to inversion problems in geophysics (Tarantola et al. (1982))

5. Exploration of the posterior distribution with important sampling

In the previous section, we indicated how the GLLiM posterior could be used for prediction. In this section, we leverage our knowledge of the true model F to enhance the predictions using importance sampling (IS). More specifically, since the real posterior $p_0(\mathbf{x} \mid \mathbf{y}) \propto \mathcal{L}_{\mathbf{x}}(\mathbf{y})\ p_0(\mathbf{x})$ is only known up to a constant, we use self-normalized importance sampling (Robert and Casella, 2004).

5.1. Mean prediction with importance sampling

The so called self-normalized importance sampling is based on the observation that, using the tractable part $\tilde{p}_0(\mathbf{x} \mid \mathbf{y}) = \mathcal{L}_{\mathbf{x}}(\mathbf{y})\ p_0(\mathbf{x})$, the posterior mean writes

$$E[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}] = \frac{\int \mathbf{x}\ \tilde{p}_0(\mathbf{x}|\mathbf{y})\ d\mathbf{x}}{\int \tilde{p}_0(\mathbf{x}|\mathbf{y})\ d\mathbf{x}} = \frac{\int \mathbf{x}\ \frac{\tilde{p}_0(\mathbf{x}|\mathbf{y})}{\nu(\mathbf{x})}\ \nu(x)\ d\mathbf{x}}{\int \frac{\tilde{p}_0(\mathbf{x}|\mathbf{y})}{\nu(\mathbf{x})}\ \nu(x)\ d\mathbf{x}}$$

for any distribution $\nu(x)$ satisfying $\nu(x) > 0$ where $\tilde{p}_0(\mathbf{x}|\mathbf{y}) > 0$. When ν is easy to compute and to simulate from, the law of large numbers justifies the approximation of $E[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}]$ by $\bar{\mathbf{x}}_{IS}(\mathbf{y}) = (\sum_{i=1}^{I} w_i)^{-1} \sum_{i=1}^{I} w_i \mathbf{x}_i$, where the $\mathbf{x}_i$'s are I *i.i.d.* realizations simulated according to ν and the w_i's are weights computed as $w_i = \tilde{p}_0(\mathbf{x}_i|\mathbf{y})/\nu(\mathbf{x}_i)$.

A natural candidate for ν is the approximate posterior provided by GLLiM $p_G(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta}^*)$ from which it is easy to simulate. For $i = 1 : I$, let $\mathbf{x}_i^G$ be simulated from p_G with the associated importance weight $w_i^G = \tilde{p}_0(\mathbf{x}_i^G|\mathbf{y})/p_G(\mathbf{x}_i^G|\mathbf{y}; \boldsymbol{\theta}^*)$. Eventually the importance sampling (IS) approximation of the mean is given by

$$\bar{\mathbf{x}}_{IS-G}(\mathbf{y}) \quad = \quad \frac{\sum_{i=1}^{I} w_i^G \mathbf{x}_i^G}{\sum_{i=1}^{I} w_i^G}\ . \tag{6}$$

5.2. Centroid-based prediction with importance sampling

Similarly, predictions with the posterior modes as introduced in Section 4 can be refined using again importance sampling. A general property is that if the prior is chosen as importance distribution,

the weights are proportional to the likelihood. Thus, one first draws samples from the prior and then determines their weights by comparing them with the data. The higher the weight, the closer to the mode provided the drawn samples lie in a region where the posterior probability is high. If this is the case, the importance sampling mean is likely to provide a solution with higher posterior probability.

To choose an appropriate prior, we propose to use the knowledge learned from the merged posterior. The prior is set to $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1^y, \boldsymbol{\Sigma}_1^y)$ (resp. $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_2^y, \boldsymbol{\Sigma}_2^y)$) to estimate the first mode (resp. the second) in cases where we focus on two modes. A more general case can be easily considered. We therefore simulate: for $i = 1 : I, \quad \mathbf{x}_i^1 \sim \mathcal{N}(\boldsymbol{\mu}_1^y, \boldsymbol{\Sigma}_1^y)$ and compute the associated importance weights $w_i^1 = \mathcal{N}(F(\mathbf{x}_i^1); \mathbf{y}, \boldsymbol{\Sigma})$. Eventually the IS approximation of the first mode is given by

$$\bar{\mathbf{x}}_{IS-centroid,1}(\mathbf{y}) \quad = \quad \frac{\sum_{i=1}^I w_i^1 \mathbf{x}_i^1}{\sum_{i=1}^I w_i^1} \, . \tag{7}$$

Finally this approach provides a potential additional refinement for the values of $\mathbf{x}$ that can be checked by calculating the reconstruction error both for a unimodal or multimodal case. More details are given in the experiments section 7.

6. Estimation of the uncertainty on the model via EM

We propose a procedure which allows to handle the situations where the uncertainty on the model is not known. This uncertainty may include both deviations from the theoretical model and measurements errors. However, estimation is possible only under the restrictive assumption that this uncertainty is represented by the same covariance matrix for all observations. In other words, the observed $\{\mathbf{y}_m, m = 1 : N_{obs}\}$ to be inverted are assumed to be independent realizations of the same model given by $\mathbf{Y} = F(\mathbf{X}) + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\mathbf{X} \sim p(\mathbf{x})$. We use the same generic notation $\boldsymbol{\Sigma}$ as in Section 3.1 although it may also include a constant $\boldsymbol{\Sigma}_{obs}$ component. The goal is then to estimate $\boldsymbol{\Sigma}$ from the $\mathbf{y}_m$'s. If the associated $\{\mathbf{x}_m, m = 1 : N_{obs}\}$ were also observed, a straightforward maximum likelihood estimation of $\boldsymbol{\Sigma}$ would be given by, $N_{obs}^{-1} \sum_{m=1}^{N_{obs}} (\mathbf{y}_m - F(\mathbf{x}_m))(\mathbf{y}_m - F(\mathbf{x}_m))^T$. As in standard missing data settings, the idea is to treat the missing $\mathbf{x}_m$ as latent variables and to use an EM algorithm to estimate $\boldsymbol{\Sigma}$. Starting from an initial value $\boldsymbol{\Sigma}^{(0)}$, the EM algorithm consists of updating at iteration (r),

$$\begin{aligned}
\boldsymbol{\Sigma}^{(r)} &= \arg\max_{\boldsymbol{\Sigma}} \sum_{m=1}^{N_{obs}} \mathbb{E}[\log p(\mathbf{y}_m, \mathbf{X}_m; \boldsymbol{\Sigma}) \mid \mathbf{y}_m; \boldsymbol{\Sigma}^{(r-1)}] \\
&= \arg\max_{\boldsymbol{\Sigma}} N_{obs} \log |\boldsymbol{\Sigma}| + \sum_{m=1}^{N_{obs}} \mathbb{E}[(\mathbf{y}_m - F(\mathbf{X}_m))^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_m - F(\mathbf{X}_m)) \mid \mathbf{y}_m; \boldsymbol{\Sigma}^{(r-1)}]
\end{aligned}$$

from which we can derive that

$$\boldsymbol{\Sigma}^{(r)} \quad = \quad \frac{1}{N_{obs}} \sum_{m=1}^{N_{obs}} \mathbb{E}[(\mathbf{y}_m - F(\mathbf{X}_m))(\mathbf{y}_m - F(\mathbf{X}_m))^T \mid \mathbf{y}_m; \boldsymbol{\Sigma}^{(r-1)}] \tag{8}$$

The expression above requires expectations with respect to $p(\mathbf{x} \mid \mathbf{Y} = \mathbf{y}_m; \boldsymbol{\Sigma}^{(r-1)})$ which we propose to compute using importance sampling with $p_G(\mathbf{x} \mid \mathbf{Y} = \mathbf{y}_m; \boldsymbol{\theta}^{(r-1)})$ as importance distribution.

But the use of the GLLiM approximation p_G requires a dictionary $\mathcal{D}^{(r-1)}$ and a learning step of the GLLiM parameters which is generally different at each iteration. To avoid high computational cost, we can however make use of the GLLiM model property already mentioned on Section 4.2 to update p_G directly by modifying the $\mathbf{\Sigma}_k$ parameters from an initial learned GLLiM model corresponding to $\boldsymbol{\theta}^{\text{ref}}$. The value of $\mathbf{\Sigma}$ used to initialize the EM algorithm is application dependent.

Algorithm 1: EM uncertainty estimation

Result: Estimated $\mathbf{\Sigma}$.
Generate a training dictionary $\mathcal{D}^{\text{ref}}$ and learn a GLLiM model $\boldsymbol{\theta}^{\text{ref}}$.
Initialize $\mathbf{\Sigma}$ to $\mathbf{\Sigma}^{(0)}$ (application dependent).
while $r \leq r_{max}$ **do**
  Update the GLLiM parameter $\boldsymbol{\theta}^{(r-1)}$ using $\boldsymbol{\theta}^{\text{ref}}$ and $\mathbf{\Sigma}^{(r-1)}$ as described in Section 4.2.
  foreach $\mathbf{y}_m$ **do** Importance Sampling
    For $i = 1 : I_m$, sample $\mathbf{x}_i$ from $p_G(\mathbf{x} \mid \mathbf{Y} = \mathbf{y}_m; \boldsymbol{\theta}^{(r-1)})$;
    Compute the weights $w_i^{(r)} = p(\mathbf{x}_i \mid \mathbf{Y} = \mathbf{y}_m; \mathbf{\Sigma}^{(r-1)}) / p_G(\mathbf{x}_i \mid \mathbf{Y} = \mathbf{y}_m; \boldsymbol{\theta}^{(r-1)})$
    Compute $S_m^{(r)} := (\sum_{i=1}^{I_m} w_i^{(r)})^{-1} \sum_{i}^{I_m} w_i^{(r)} (\mathbf{y}_m - F(\mathbf{x}_i))(\mathbf{y}_m - F(\mathbf{x}_i))^T$
  end
  update $\mathbf{\Sigma}^{(r-1)}$ according to (8) as $\mathbf{\Sigma}^{(r)} = N_{obs}^{-1} \sum_{m=1}^{N_{obs}} S_m^{(r)}$
end

7. Illustrations

The different prediction schemes introduced earlier are tested on various experiments. The first part of this section is devoted to toy examples, designed to illustrate potential numerical issues for which we propose solutions. The second part tackles our main real data application: an inverse problem in remote sensing. The algorithms are currently implemented in the Julia programming language (Bezanson et al., 2017), and available online. Computation times are provided in Appendix A. The notations previously introduced are recalled in Table 1.

Notation	Description and dimension
$\mathbf{y}_{obs}$	observation (dimension D)
$\mathbf{x}_{obs}$	for synthetic experiments, original true parameters (dimension L)
$\mathbf{x}_{obs,1}$	for synthetic multi solution experiments, first original true parameters
$\mathbf{x}_{obs,2}$	for synthetic multi solution experiments, second original true parameters
$\bar{\mathbf{x}}_G$	mean of the GLLiM posterior density (dimension L)
$\hat{\mathbf{x}}_{centroid,1}$	first centroid of the merged GLLiM density (dimension L)
$\hat{\mathbf{x}}_{centroid,2}$	second centroid of the merged GLLiM density (dimension L)
$\bar{\mathbf{x}}_{IS-G}$	Importance Sampling estimator of the mean of the posterior (dimension L)
$\bar{\mathbf{x}}_{IS-centroid,1}$	Importance Sampling estimator of the first mode of the posterior (dimension L)
$\bar{\mathbf{x}}_{IS-centroid,2}$	Importance Sampling estimator of the second mode of the posterior (dimension L)
$\hat{\mathbf{x}}_{best}$	best prediction among all estimators (with respect to the reconstruction error)

TABLE 1
Notations

	D	L	K	N	σ	N_{obs}	α_{obs}	I
Example 1	1	1	40	50000	0.01	100	1000	10000
Example 2	9	4	30,40,70	50000	0.001	1000	1000	10000
Example 3	10	4	40,70,100	50000	0.0001	1000	1000	50000
Example 4	10	4	50	50000	0.001	100	20	50000
Example 5	11	4	50	50000	0.001	154650	NA	20000

TABLE 2
Synthetic and real data examples. Training and testing settings.

7.1. Validation on simple models and synthetic examples

7.1.1. Procedure

A similar procedure is adopted for all the tests in this part. In most applications, for a given forward model F, the training data set is simulated using a range of values for each parameter. Without loss of generality, the parameter space is then assumed to be $\mathcal{P} = [0,1]^L$ after normalization of each parameter if necessary. A set of N training parameter vectors is generated randomly using Sobol quasi-random sequences (Sobol, 1967) to get evenly spaced samples in $\mathcal{P}$. The corresponding N training observations ($\mathbf{y}_n$'s) are then generated by applying the forward model to each simulated parameter vector and adding some centered Gaussian noise with covariance matrix $\mathbf{\Sigma} = \sigma^2 Id$. It follows a set $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n), n = 1 : N\}$ with $\mathbf{y}_n = F(\mathbf{x}_n) + \boldsymbol{\epsilon}_n$ with $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. $\mathcal{D}$ is used to learn a GLLiM model with K mixture components as described in Section 3. A test set $\mathcal{T} = \{(\mathbf{x}^m_{obs}, \mathbf{y}^m_{obs}), m = 1 : N_{obs}\}$ is simulated similarly, with N_{obs} parameter vectors $\mathbf{x}^m_{obs}$ and $\mathbf{y}^m_{obs} = F(\mathbf{x}^m_{obs})$. To mimic standard remote sensing situations, this test set is supposed to come with additional information on the measurement uncertainty. Each observation $\mathbf{y}^m_{obs}$ is assumed to be corrupted with a centered Gaussian noise with covariance matrix $\mathbf{\Sigma}^m_{obs}$. Following common practice in remote sensing (*e.g.* Schmidt and Fernando (2015)), $\mathbf{\Sigma}^m_{obs}$ is generally taken as $\mathbf{\Sigma}^m_{obs} = diag(F(\mathbf{x}^m_{obs})/\alpha_{obs})^2$ for some positive scalar α_{obs} to be specified representing a percentage of uncertainty.

For predictions, the different schemes are compared leading to several possible values $\mathbf{x}$ for each observation $\mathbf{y}_{obs} \in \mathcal{T}$ in the test set. More specifically, we consider up to 6 different predictions all based on the initial GLLiM model with K components. As explained in Section 4.2, the learned GLLiM parameters are adjusted to account for $\mathbf{\Sigma}^n_{obs}$'s. Then the 6 prediction values are: the GLLiM expectation value ($\bar{\mathbf{x}}_G$), the two GLLiM mixture centroids ($\hat{\mathbf{x}}_{centroid,1}$ and $\hat{\mathbf{x}}_{centroid,2}$), and their counterparts obtained after importance sampling ($\bar{\mathbf{x}}_{IS-G}, \bar{\mathbf{x}}_{IS-centroid,1}, \bar{\mathbf{x}}_{IS-centroid,2}$). The importance sampling procedure requires the simulation of I parameter vectors according to the chosen proposal distribution. The specific values of $D, L, K, N, \sigma, N_{obs}, \alpha_{obs}, I$ are specified for each experiment and recap in Table 2.

The predictions are compared using the (relative) reconstruction error (sometimes also called residual error) defined as $R(\mathbf{x}) = ||\mathbf{y}_{obs} - F(\mathbf{x})||_2 / ||\mathbf{y}_{obs}||_2$. In the following tables, $\bar{R}$ will denotes the reconstruction error averaged over the test dataset. When ground truth values $\mathbf{x}_{obs}$ of the parameters are available, another performance criterion is the prediction error on $\mathbf{x}$ denoted by $E(\mathbf{x}) = ||\mathbf{x} - \mathbf{x}_{obs}||_\infty$. Similarly $\bar{E}$ will refer to the error averaged over the test dataset. When the parameters space is $[0,1]^L$ this quantity belongs to $[0,1]$ making it comparable from one experience to another.

In practice, we suggest to retain the value, denoted by $\mathbf{x}_{best}$, that minimizes the reconstruction error among all the proposed estimators.

7.1.2. Example 1: Simple double solution problem

A first example illustrating the multiple solutions case is obtained with $F(x) = (x-0.5)^2$ ($L = D = 1$), which admits two solutions denoted by $\mathbf{x}_{obs,1}$ and $\mathbf{x}_{obs,2} = 1 - \mathbf{x}_{obs,1}$ for every observation. In this example, $N = 50000, \sigma = 0.01$ and $K = 40$. For the test set, $N_{obs} = 100$ parameter values are sampled along sinus functions for visualization purpose and observations are generated applying F with a noise set with $\alpha_{obs} = 1000$. $I = 10000$ samples are drawn to perform importance sampling. Figure 1 shows that predictions based on the mean cannot retrieve the true solutions, while centroid predictions after mixture merging are accurate. The IS-GLLiM mean is close to the GLLiM mean showing that the inaccuracy is not coming from the GLLiM approximation.

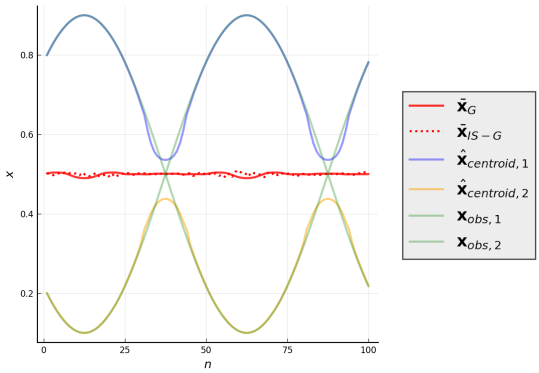

FIG 1. *Double solution case. Mean predictions using GLLiM and refined with importance sampling (red) vs centroid predictions (blue and yellow) using the two-component mixture obtained by merging the GLLiM mixture components. The true parameter values* $\mathbf{x}_{obs,1}$ *and* $\mathbf{x}_{obs,2}$*, chosen along sinus functions, are in green.*

7.1.3. Example 2: Importance sampling for the centroids

In this example, the goal is to illustrate how importance sampling may correct the imprecision due to the approximation induced by the first GLLiM mapping step. F is designed so as to exhibit 2 solutions with $D = 9$ and $L = 4$, $F = A \circ G \circ H$, where A is a $D \times L$ injective matrix, $G(\mathbf{x}) = (\exp(x_1), \exp(x_2), \exp(x_3), \exp(x_4))$ and $H(\mathbf{x}) = (x_1, x_2, 4(x_3 - 0.5)^2, x_4)$. The resulting F is therefore non-linear and yields two solutions for each observation, denoted by $\mathbf{x}_{obs,1}$ and $\mathbf{x}_{obs,2} = 1 - \mathbf{x}_{obs,1}$. In this example, $N = 50000, \sigma = 0.001, N_{obs} = 1000, \alpha_{obs} = 1000$. Three different values of $K = 30, 50, 70$ are tested. As in previous simulations, the N_{obs} $\mathbf{x}_{obs}^n$'s are set along a sinus function. Figure 2 shows that importance sampling significantly improves predictions. This is confirmed by the reconstruction and prediction errors reported in Table 3. To handle the double solutions in a meaningful way we compute these errors by comparing the pair $(\mathbf{x}_{obs,1}, \mathbf{x}_{obs,2})$ with the most favorable permutation of the two predicted centroids. In particular, Table 3 shows that K has not a huge impact on the results and that the quality of the GLLiM approximation is not the major factor in the IS prediction quality.

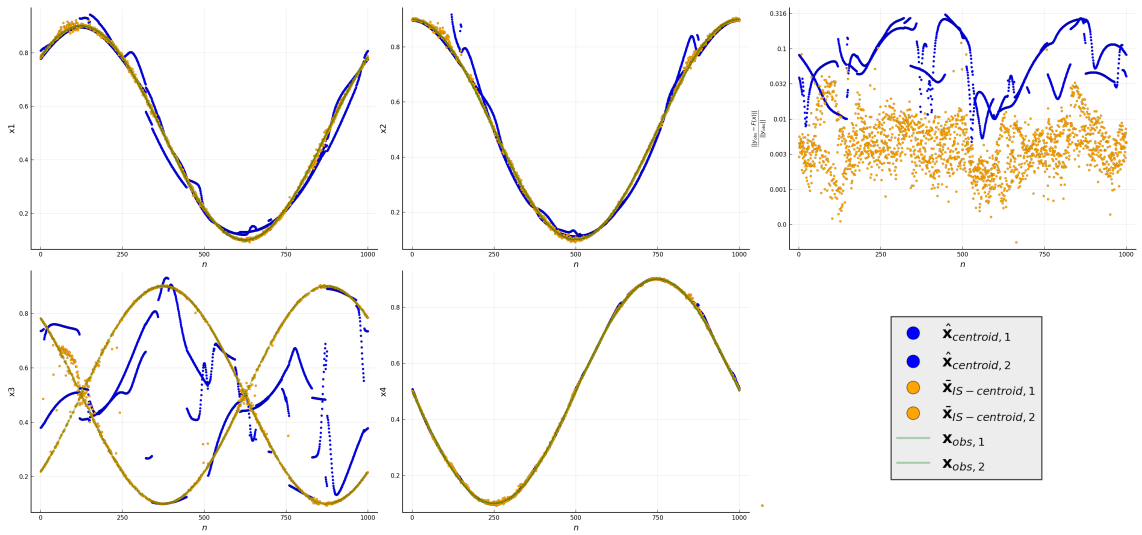

FIG 2. *Importance sampling for centroids. IS centroid predictions (yellow) starting from GLLiM with $K=30$ (blue). The $L=4$ parameters are shown separately with the true centroids in green. Reconstruction errors are in the top right plot.*

Prediction scheme	K = 30	K = 50	K = 70
$\bar{R}(\hat{\mathbf{x}}_{centroid,1})$	0.0956 (0.0654)	0.0821 (0.0699)	0.0776 (0.0767)
$\bar{R}(\hat{\mathbf{x}}_{centroid,2})$	0.0711 (0.0512)	0.0681 (0.0469)	0.078 (0.0644)
$\bar{R}(\bar{\mathbf{x}}_{IS-centroid,1})$	**0.0079** (0.011)	0.0076 (0.0109)	**0.0062** (0.0135)
$\bar{R}(\bar{\mathbf{x}}_{IS-centroid,2})$	0.0094 (0.0125)	**0.006** (0.0088)	0.009 (0.0212)
$\bar{E}(\hat{\mathbf{x}}_{centroid,1})$	0.1224 (0.1108)	0.1073 (0.0748)	0.1104 (0.0863)
$\bar{E}(\hat{\mathbf{x}}_{centroid,2})$	0.142 (0.0964)	0.1324 (0.1161)	0.1214 (0.0865)
$\bar{E}(\bar{\mathbf{x}}_{IS-centroid,1})$	**0.0504** (0.1401)	**0.058** (0.1375)	**0.0488** (0.1436)
$\bar{E}(\bar{\mathbf{x}}_{IS-centroid,2})$	0.0727 (0.1715)	0.0594 (0.1575)	0.0618 (0.1468)

TABLE 3

Importance sampling for centroids. Average reconstruction (first 4 lines) and prediction (last 4 lines) errors, for 4 prediction schemes, 1000 tests, 3 GLLiM settings. Standard deviations are in parenthesis and best averages are in bold.

7.2. A physical model inversion in planetary science

Our real data application comes from the study of the Martian environment, in particular the morphological, compositional and textural characterization of sites representing various geological contexts at different periods of the history of Mars. The composition of the materials is established on the basis of spectral mixing and physical modelling techniques using images produced by hyperspectral cameras (Compact Reconnaissance Imaging Spectrometer for Mars (Murchie et al., 2009) or CRISM@MRO). Information on the microtexture of surface materials such as grain size, shape, roughness and internal structure can also be used as tracers of geological processes (Fernando, Schmidt and Douté, 2016). This information is accessible under certain conditions thanks to hyperspectral image sequences acquired from eleven different angles by the CRISM instrument during a site flyover by MRO. It should be noted that such observations can also be measured in the laboratory, on known materials in order to validate a model, or on materials of unknown origin,

such as meteorite fragments (Potin et al., 2019). In both cases, the interpretation of the surface Bidirectional Reflectance Distribution Factor (BRDF) extracted from these observations, in terms of composition and microtexture, is based on the inversion of physical models of radiative transfer linking physical and observable parameters in a non-linear way.

The Hapke model is a semi-empirical photometric model that relates physically meaningful parameters to the reflectivity of a granular material for a given geometry of illumination and viewing. Formally, it links a set of parameters $\mathbf{x} \in \mathbb{R}^4$ to a *theoretical* BRDF denoted by $\mathbf{y} = F_{hapke}(\mathbf{x}) \in \mathbb{R}^D$. A given experiment defines D geometries of measurement, each parametrized by a triplet (θ_0, θ, ϕ) of incidence, emergence and azimuth angles. Moreover, $\mathbf{x} = (\omega, \bar{\theta}, b, c)$ are the sensitive parameters (respectively single scattering albedo, macroscopic roughness, asymmetry parameter and backscattering fraction). For simplicity, we ignore the angular width h and the amplitude B_0 of the opposition effect. More details on these notations and their photometric meaning may be found for example in Schmidt and Fernando (2015), alongside the explicit expression of F_{hapke}.

In the following experiments, we also apply a change of variable $\gamma = 1 - \sqrt{1-\omega}$, in order to avoid a high non-linearity of F_{hapke} when ω tends to 1. Indeed F_{hapke} is equivalent to $\sqrt{1-\omega}$ when ω is close to 1. This may lead to infinite derivatives with respect to ω, a challenging situation for a locally linear model. Better results are observed for some edge case observations while the change does not impact other cases.

The experiments in this section are all made using similar learning sets for the initial GLLiM model. The number of parameters is $L = 4$ with $D = 10$ or 11 geometries considered. The size of the training set is $N = 50000$ and different values of K are considered as summarized in Table 2. Predictions are obtained following the same procedure as in the previous section.

7.2.1. Example 3: Synthetic data from the Hapke's model

Prior to real data inversion, a first step is to check the ability of our method to accurately capture the Hapke's model main features. To this end, $N_{obs} = 1000$ synthetic observations are generated. The $L = 4$ parameters above are considered and each of them is simulated along a sinus function for visualization purpose. For $m = 1 : N_{obs}$ and $l = 1 : L$, $\mathbf{x}_{obs}^{m,l} = 0.5 + 0.4\sin(\frac{2m\pi}{N_{obs}} + \frac{l\pi}{4})$. The corresponding reflectance curves are generated as $\mathbf{y}_{obs}^m = F_{hapke}(\mathbf{x}_{obs}^m)$ corrupted by a noise defined by $\alpha_{obs} = 1000$. The dimension of the observed vector ($D = 10$) and the measurement geometries used to define F_{hapke} are borrowed from a real laboratory experiment presented in the next example. The experimental setting defines geometries at which the measurements are made, which in turn define F_{hapke}. The number of geometries thus corresponds to the size D of each observation. In this experiment $D = 10$. All the prediction schemes are performed but only the GLLiM expectation prediction $\bar{\mathbf{x}}_G$, its refinement by importance sampling $\bar{\mathbf{x}}_{IS-G}$ and the best prediction in terms of reconstruction error, denoted by $\hat{\mathbf{x}}_{best}$ are reported in Figure 3. This example highlights the interest of the different predictions methods. The GLLiM expectation prediction $\bar{\mathbf{x}}_G$ is often good enough, the centroid estimations not giving additional information, apart from confirming the uni-modality of the posterior. See the estimation of ω in Figure 3 where all predictions coincide. However, it appears on the other parameters that a better precision can be reached using importance sampling starting from the GLLiM posterior approximation as proposal distribution. This is visible in the reconstruction error plot of Figure 3. At last, in some cases, prediction by the mean can be further improved by considering centroids. This corresponds to the case where the $\hat{\mathbf{x}}_{best}$ estimation significantly differs from the two others. For a more quantitative analysis, Table 4

provides the reconstruction and prediction errors for these observations. The centroid predictions, not plotted for clarity, are added in this table.

Prediction scheme	K = 40	K = 70	K = 100
$\bar{R}(\bar{\mathbf{x}}_G)$	0.0718 (0.0862)	0.0417 (0.0337)	0.0365 (0.0443)
$\bar{R}(\bar{\mathbf{x}}_{IS-G})$	0.0049 (0.0049)	0.0034 (0.0031)	0.0037 (0.0045)
$\bar{R}(\bar{\mathbf{x}}_{IS-centroid,1})$	0.0087 (0.0188)	0.0073 (0.0109)	0.006 (0.0101)
$\bar{R}(\bar{\mathbf{x}}_{IS-centroid,2})$	0.0103 (0.0128)	0.0078 (0.01)	0.0057 (0.0093)
$\bar{R}(\hat{\mathbf{x}}_{best})$	**0.0026** (0.0022)	**0.0019** (0.0015)	**0.0018** (0.0016)
$\bar{E}(\bar{\mathbf{x}}_G)$	0.1557 (0.127)	0.1245 (0.1051)	0.1175 (0.1001)
$\bar{E}(\bar{\mathbf{x}}_{IS-G})$	0.0753 (0.091)	0.0519 (0.0698)	0.0607 (0.0807)
$\bar{E}(\bar{\mathbf{x}}_{IS-centroid,1})$	0.0699 (0.0868)	0.0629 (0.0795)	0.0698 (0.1065)
$\bar{E}(\bar{\mathbf{x}}_{IS-centroid,2})$	0.1243 (0.1544)	0.102 (0.1359)	0.0662 (0.0837)
$\bar{E}(\hat{\mathbf{x}}_{best})$	**0.0398** (0.0457)	**0.031** (0.0367)	**0.033** (0.0411)

TABLE 4

Synthetic data from the Hapke's model. Average reconstruction (first 4 lines) and prediction (last 4 lines) errors, for 4 prediction schemes, 1000 tests, 3 GLLiM settings. Standard deviations are in parenthesis and best averages are in bold.

7.2.2. Example 4: Laboratory observations

Reflectance measurements in the laboratory on crushed minerals are now considered (see Pilorget et al. (2016) for more details). We focus on three experiments (Olivine, Nontronite and Basalt), each of them measured at 100 wavelengths in the spectral range 400-2800 nm, seen as independent observations. For our test we select the Nontronite dataset since it presents the largest spectral variations. It corresponds therefore to a number of $N_{obs} = 100$ observations $\mathbf{y}_{obs}^m$ $(m = 1 : N_{obs})$ to invert. As before, the experimental setting defines geometries at which the measurements are made, which in turn define F_{hapke}. The size D of each observation is $D = 10$ and the corresponding angles are such that the incidence and azimuth angles are fixed to $\theta_0 = 45$ and $\Phi = 0$. In Figures 4 and 5, our results are compared to that obtained with MCMC techniques, used for example in Schmidt and Fernando (2015). For the MCMC procedure, a Metropolis-Hasting algorithm was used, generating for each of the 100 observations to be inverted 10^7 samples with a burnin period of 5×10^6 samples. For a comparison with Schmidt and Fernando (2015), the noise added to the observations $\mathbf{y}_{obs}^m$'s is, in this example, $\mathbf{\Sigma}_{obs}^m = diag(\max(\mathbf{y}_{obs}^m/20, 0.01)^2)$. This corresponds most of the time to a relative error of 5%. For very dark material, the reflectance $\mathbf{y}_{obs}^m$ is low. The above formula therefore ensures a minimum noise of 0.01 unit of reflectance in all cases. Figure 4 provides the inversion results for the Nontronite measurements for each wavelength, in the form of spectra for each parameter and for the reconstruction error. The parameter ω is accurately retrieved. The three other parameters show abrupt variations between successive observations. This is not physically satisfying and probably results from the relative low sensitivity of the direct model to these parameters, yielding posterior distributions with no clear peaks, responsible in turn for rather unstable predictions in the $\mathcal{X}$ space. Figure 5 compares the observation to reconstructions from predictions. For this last figure, the results are shown via polar plots, the angle being the emergence angle and the radius the level of reflectance. The yellow line indicates the illumination direction. Three different wavelengths among the $\mathbf{y}_{obs}^m$'s are chosen to fall on specific spectral features. The plots show then 10 points which coordinates are $(\mathbf{y}_{obs}^{m,d}, \theta_d)$ or $(F_{hapke}(\mathbf{x})_d, \theta_d)$ for $d = 1 : D$ and for different predictions $\mathbf{x}$. The MCMC prediction targets the mean and is therefore closer to our mean predictions while it appears that our $\hat{\mathbf{x}}_{best}$ predictions provide better reconstructions.

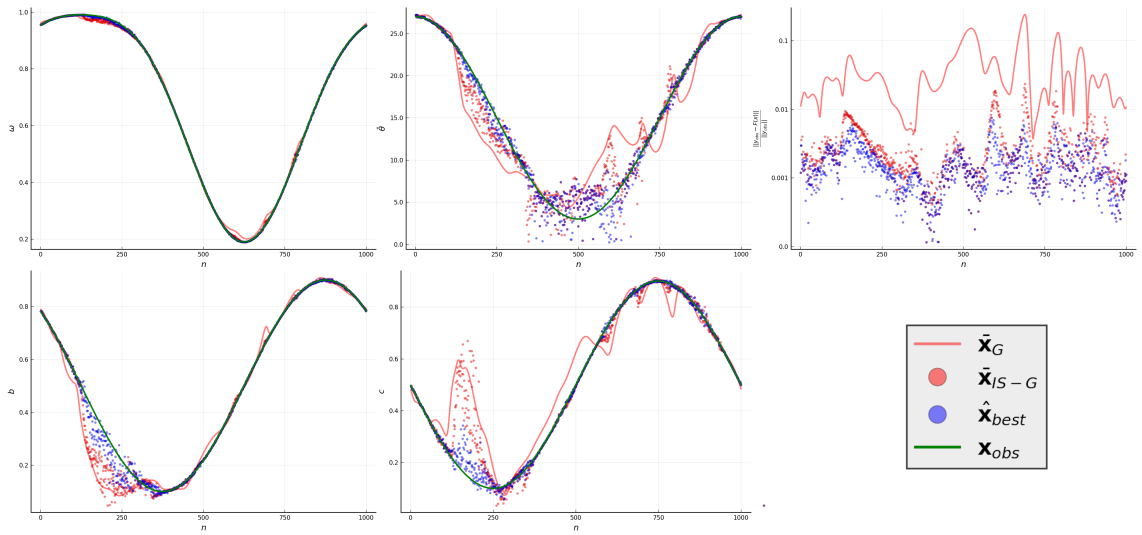

FIG 3. *Inversion of synthetic data from the Hapke's model. A GLLiM model is learned from training data with* $K = 70$. *The GLLiM mean prediction* $\bar{\mathbf{x}}_G$ *(plain, red), its refined version with importance sampling* $\bar{\mathbf{x}}_{IS-G}$ *(circle, red) and the reconstruction best prediction* $\hat{\mathbf{x}}_{best}$ *(circle, blue), among all prediction schemes including the centroids, are compared. The values to be recovered* $\mathbf{x}_{obs}$ *are in green.*

7.2.3. Uncertainty level estimation

In this section, the algorithm 1 described in Section 6 to assess the uncertainty in the model is tested. Synthetic data are first considered with the same setting, F_{hapke}, $D = 10, L = 4$, as in the inversion of Nontronite observations. Observations are simulated with $\Sigma = \sigma^2 I_D$ and $\sigma = 0.2$ (high uncertainty level, Figure 6 (a)) or $\sigma = 0.03$ (low uncertainty level, Figure 6 (b)), satisfying the assumptions made in Section 6 and Algorithm 1. To simplify the notation, we will denote by $\boldsymbol{\sigma}_D$ the diagonal matrix containing the D standard deviations here all equal to σ. The EM algorithm is constrained to provide diagonal covariances. It is initialized with $\boldsymbol{\Sigma}^{(0)} = 0.5 I_D$. The choice of $\boldsymbol{\Sigma}^{(0)}$ is application dependent and should be of the magnitude of the observations.

Figure 6 shows the evolution of the distance between the estimated diagonal $\boldsymbol{\sigma}_D^{(r)}$ at iteration r and the true $\boldsymbol{\sigma}_D$. For comparison, the distance between the true $\boldsymbol{\sigma}_D$ and the maximum likelihood estimator $\boldsymbol{\sigma}_D^{est}$ defined in Section 6 and calculated with the true $\mathbf{x}$'s is also provided. The figures show in addition the final value for $\boldsymbol{\sigma}_D^{(r)}$ denoted by $\boldsymbol{\sigma}_D^{last}$. These simulations illustrate that the algorithm performs satisfactorily with moderate to high uncertainty levels, but is relatively less efficient with low levels because of a lack of sensitivity to such levels.

We then consider three real datasets (Nontronite, Basalt and Olivine), again from the laboratory experiment described in Example 5. Recall that for each dataset, diagonal $\boldsymbol{\Sigma}_{obs}^m$'s accounting for measurement uncertainty are considered in addition to the measurements themselves. This information is provided by experts and can be used as a reference to assess the quality of our estimation. In this case the assumptions of Section 6 and Algorithm 1 do not hold as the uncertainty cannot be represented as a constant $\boldsymbol{\Sigma}$ across observations. Our estimations are then compared to $\boldsymbol{\sigma}_D^{obs}$ set to the average over N_{obs} of the square roots of the $\boldsymbol{\Sigma}_{obs}^m$'s which are themselves of the order of

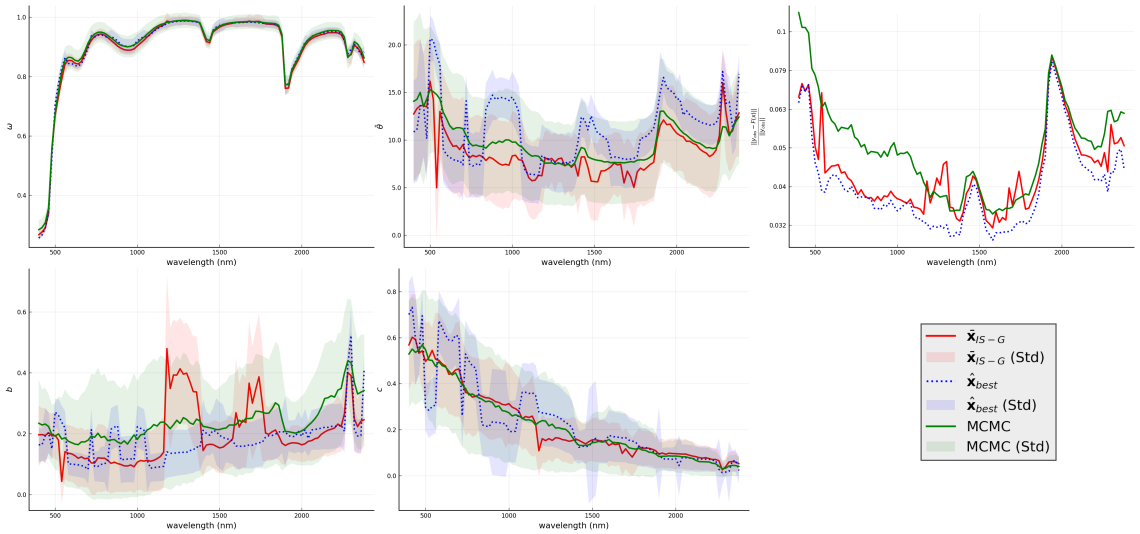

FIG 4. *Inversion of Nontronite laboratory observations. A GLLiM model is learned from training data with $K = 100$. The GLLiM mean prediction refined with importance sampling $\tilde{\mathbf{x}}_{IS-G}$ (red), the reconstruction best prediction $\hat{\mathbf{x}}_{best}$ (blue) and the MCMC prediction (green) are compared. Standard deviations are also shown.*

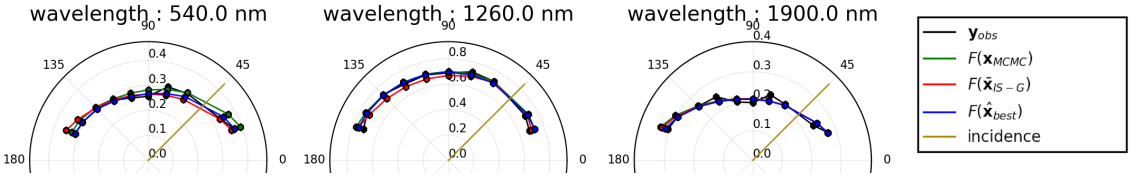

FIG 5. *Nontronite observation reconstructions. Comparison between original (black) and reconstructed observations (green, red and blue) along the emergence angle, for three wavelengths : 540, 1260 and 1900 nm. The incidence and azimuth angles are equal to 45 and 0 respectively.*

$\mathbf{\Sigma}^m_{obs} = diag(\mathbf{y}^m_{obs}/20)^2$. In addition, it is usually expected that the standard deviation is not too much dependent on the geometry. Following this expectation, the EM algorithm is run with an isotropic constraint on $\mathbf{\Sigma}$.

Figure 7 shows the estimated uncertainty level (more specifically, the standard deviations for each geometry) using polar plots. It appears that the estimated standard deviations are of the same order as those given by the experts in particular for Olivine and Nontronite data sets. When expert references are considered as reliable, the observed deviations may be interpreted as an indication of the lesser ability of the theoretical model F_{hapke} to provide a good modelling of the observations. However, for the Basalt results, the deviation is more likely to come from an underestimation of the expert measurement error due to the darkness of the material.

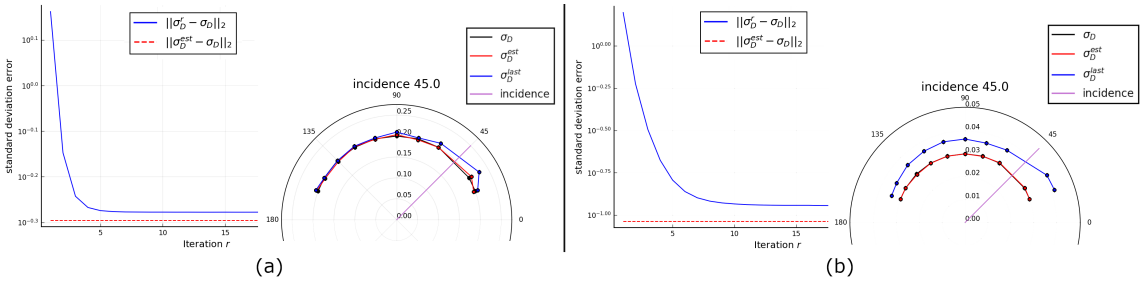

FIG 6. *Noise estimation on synthetic data. True value is* $\sigma = 0.2$ *(a) and* $\sigma = 0.03$ *(b). Left: evolution of* $||\boldsymbol{\sigma}_D^{(r)} - \boldsymbol{\sigma}_D||_2$ *(blue) and* $||\boldsymbol{\sigma}_D^{est} - \boldsymbol{\sigma}_D||_2$ *(orange). Right: comparison between final value* $\boldsymbol{\sigma}_D^{last}$ *(blue),* $\boldsymbol{\sigma}_D^{est}$ *(red) and* $\boldsymbol{\sigma}_D$ *(black).*

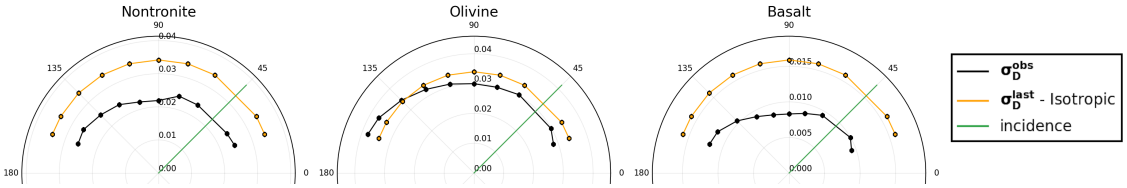

FIG 7. *Noise level estimation on laboratory data. Comparison between the given* $\boldsymbol{\sigma}_D^{obs}$ *(black) and the estimated* $\boldsymbol{\sigma}_D^{last}$ *provided by the EM algorithm (isotropic constraint, in orange), along geometries of measurements.*

7.2.4. Example 5: Massive inversion of spatial and spectral Mars data

We conclude our experiments with a large scale inversion. The dataset comes from a multi-angular observation of the South Pole of Mars by the Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) (Murchie et al., 2009). The targeted scene presents spatially segregated C02 ice, H2O frost, and mineral dust (Douté and Pilorget, 2017). After fusion and atmospheric correction of eleven hyperspectral images (Ceamanos et al., 2013), the dataset provides both spatial and spectral dimensions, totaling $N_{obs} = 154650 = 3093 \times 50$ measurements vectors $\mathbf{y}_{obs}^m$, which makes MCMC approaches unacceptably slow. Each $\mathbf{y}_{obs}^m$ is provided with an additional diagonal $\mathbf{\Sigma}_{obs}^m$ accounting for measurement uncertainty. Maps of Hapke's parameter values are generated from the results of our inversion and superposed with transparency onto the full resolution CRISM nadir image, which serves as a geological control background image. The maps are shown in Figure 8 for ω and $\bar{\theta}$ and in Appendix B for b and c. The methods with and without taking into account the centroids ($\hat{\mathbf{x}}_{best}$ and $\bar{\mathbf{x}}_{IS-G}$ respectively) are compared. The results are satisfying from the application point of view. The colour composition of Figure 8 first line reflects the variation of ω at three wavelengths and corresponds well with the spatial distribution of the three previous materials and their known spectral optical properties. The map of $\bar{\theta}$ averaged over the spectral dimension is color coded by intervals of values whose spatial variations are correlated with the composition and the structures of the terrains. In general (see also maps b and c in Appendix B), all predicted parameters preserve some spatial regularity and show meaningful correlations with the composition and the geology. Moreover, the prediction including centroids ($\hat{\mathbf{x}}_{best}$) brings some improvement in the experiment: it mostly agrees with the mean prediction ($\bar{\mathbf{x}}_{IS-G}$) while increasing the inversion success rate and the spatial regularity. For example, the first line of Figure 8 shows color artefacts that partly disappear

with $\hat{\mathbf{x}}_{best}$.

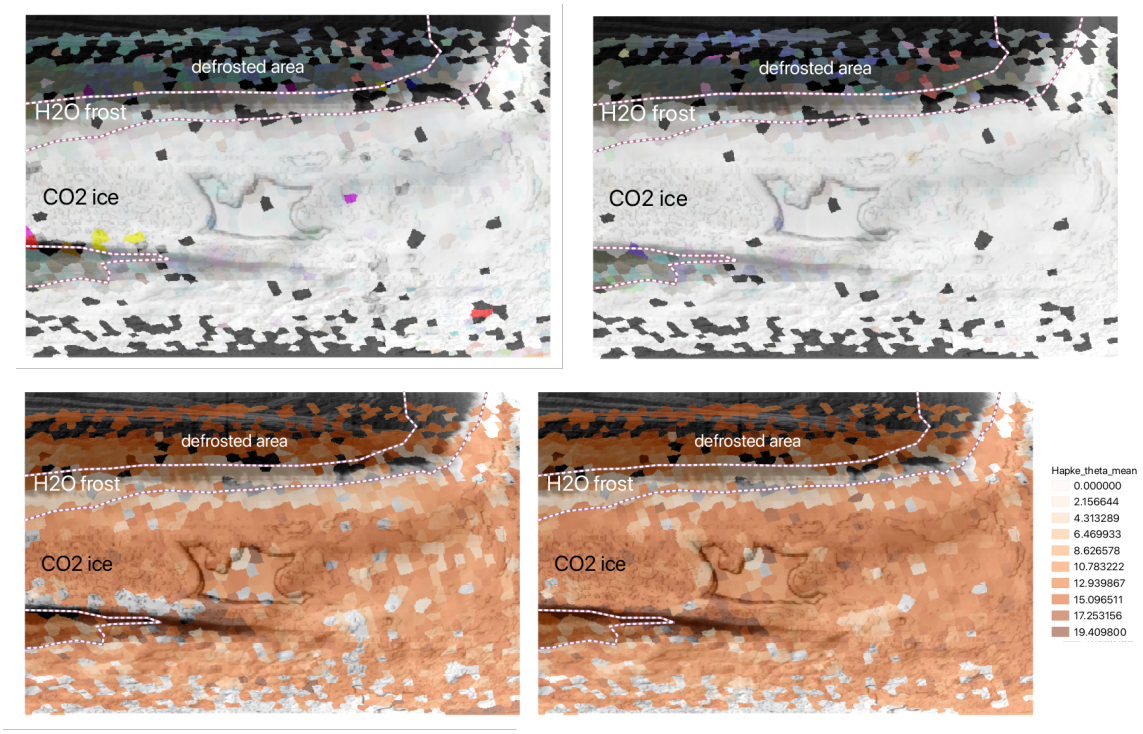

FIG 8. *Mars South Pole dataset. Parameter ω in synthetic colors coding 3 wavelengths (top) and parameter $\bar{\theta}$ averaged over spectral dimension (bottom), predicted using $\bar{\mathbf{x}}_{IS-G}$ (left) or $\hat{\mathbf{x}}_{best}$ (right). The darkest (top) or colorless (bottom) patches correspond to missing data or failed inversion by GLLiM.*

8. Conclusion and future work

This paper introduced a new statistical approach to inverse problems that we referred to as *fast Bayesian inversion*. The originality comes from the preliminary estimation of a global inversion operator in place of successive individual inversions. This operator is obtained via a so-called GLLiM inverse regression model and takes a simple parametric form, which once learned, provides at low cost approximations to the target posterior distributions. These approximate posteriors are subsequently refined using importance sampling. We showed the ability of the method to carry out Bayesian inversion in a computationally efficient way that allows to handle massive inversions while maintaining the advantages of a Bayesian analysis. It improves prediction accuracy, reduces the estimation time, while providing a rich information on parameter estimates via posterior distributions. This later information is essential in complex inverse problems when multiple equivalent solutions can exist. Using a physical model inversion issue in planetary remote sensing, we illustrated that we could obtain very satisfying results in situations where traditional MCMC approaches are not tractable.

The flexibility of the proposed approach opens the way to further improvements. Importance sampling was used only in its simplest version. There is a vast literature on importance sampling that could certainly be exploited. Sampling re-sampling or more elaborate forms of importance sampling via sequential Monte Carlo techniques could certainly lead to even more efficient procedures. Overall, importance sampling can be seen as a way to enhance GLLiM predictions but in a symmetric manner, the GLLiM posteriors can also be seen as good candidates for an informed importance proposal distribution, especially when dealing with multi-modal posteriors. Future work should also include the adaptation and optimization of the learning set with respect to the targeted range of parameters. In other words, the possibility to use more informative priors on the parameter space could be explored to take even more benefits from the Bayesian approach.

Appendix A: Computation times

The simulations ran on a laptop with 4 cores (at 2.5 Ghz). Table 5 shows computation times. For each experiment, the time is divided in two parts, the time for the learning step (GLLiM inference) and the time for the prediction step, which consists either of mixture merging, mode-finding and importance sampling (Examples 1 to 5) or of noise level estimation via the EM algorithm. Most experiments run in few minutes. The complexity of the forward model and the way it is implemented can take an important part in the resulting running time. This appears in the comparison of examples 2 and 3 which mainly differ in the choice of F. The Hapke model (Example 3) benefits from a more efficient implementation which explains running time twice smaller for similar settings. For equivalent forward model implementations, the time depends mainly on the size and dimensionality of the learning set and on the number of inversions to be performed. Learning sets have equivalent complexity in our experiments. Higher computation times are observed in case of massive inversions. In particular, Example 5 with 154,650 inversions takes few hours. We believe it is the first time that such spatial and spectral parametric maps are obtained due to the intractability of other methods in this setting.

Appendix B: Massive inversion of spatial and spectral Mars data

Figure 9 shows the maps for parameters b and c after inversion of the real Mars data described in Section 7.2.4.

Acknowledgements

This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02).

References

Adler, J. and Öktem, O. (2017). Solving Ill-Posed Inverse Problems Using Iterative Deep Neural Networks. *Inverse Problems* **33** 124007.

Arridge, S., Maass, P., Öktem, O. and Schönlieb, C.-B. (2019). Solving Inverse Problems Using Data-Driven Models. *Acta Numerica* **28** 1–174.

Experiment	L	D	K	N	I	N_{obs}	Learn.	$\bar{\mathbf{x}}_G$	Pred. - 1 obs.
Ex. 1 : Simple double solutions problem	1	1	40	$5\,10^4$	10^4	100	1m 34s	0.03s	11.2s - 0.11s
Ex. 2 : Centroids for double solutions	4	9	30	$5\,10^4$	10^4	10^3	2m 40s	0.41s	23m - 1.4s
-	4	9	50	$5\,10^4$	10^4	10^3	4m 26s	0.66s	23m - 1.4s
-	4	9	70	$5\,10^4$	10^4	10^3	6m 12s	1.1s	23m - 1.4s
Ex. 3 : Hapke - Synthetic data	4	10	40	$5\,10^4$	$5\,10^4$	10^3	3m 52s	0.48s	12m - 0.71s
-	4	10	70	$5\,10^4$	$5\,10^4$	10^3	6m 41s	0.86s	12m - 0.75s
-	4	10	100	$5\,10^4$	$5\,10^4$	10^3	9m 19s	1.2s	12m - 0.75s
Ex. 4 : Hapke - Laboratory observations - MCMC	4	10	-	-	10^7 *	100	-	-	29m - 17.6s
Ex. 4 : Hapke - Laboratory observations	4	10	100	10^5	$5\,10^4$	100	19m	0.15s	1m 22s - 0.82s
Ex. 5 : Hapke - Glace observations	4	11	50	$5\,10^4$	$2\,10^4$	154 650	5m 7s	1m 37s	8h 15m - 0.19s
Noise estimation on synthetic data - $\sigma = 0.2$	4	10	50	$5\,10^4$	$2\,10^4$	$2\,10^3$	4m 47s	-	1h 10m
Noise estimation on synthetic data - $\sigma = 0.03$	4	10	50	$5\,10^4$	$2\,10^4$	$2\,10^3$	4m 47s	-	1h 10m
Noise estimation on Nontronite	4	10	50	$5\,10^4$	10^4	100	4m 47s	-	2m 1s
Noise estimation on Basalt	4	10	50	$5\,10^4$	10^4	100	4m 48s	-	1m 57s
Noise estimation on Olivine	4	10	50	$5\,10^4$	10^4	100	4m 56s	-	1m 40s

TABLE 5

*Settings and computation times for synthetic and real data experiments. The column $\bar{\mathbf{x}}_G$ displays the time spend only for the computation of $\bar{\mathbf{x}}_G$ for all the observations, while the last column takes into account all the prediction schemes. * For MCMC, the number of simulations done for each observation is reported under I.*

BALSIGER, F., STEINDEL, C., ARN, M., WAGNER, B., GRUNDER, L., EL-KOUSSY, M., VALENZUELA, W., REYES, M. and SCHEIDEGGER, O. (2018). Segmentation of Peripheral Nerves From Magnetic Resonance Neurography: A Fully-Automatic, Deep Learning-Based Approach. *Frontiers in Neurology* **9** 777.

BARBIERI, M., BRIZI, L., GIAMPIERI, E., SOLERA, F., CASTELLANI, G., TESTA, C. and REMONDINI, D. (2018). Circumventing the Curse of Dimensionality in Magnetic Resonance Fingerprinting through a Deep Learning Approach. *arXiv:1811.11477 [physics]*.

BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards Scaling up Markov Chain Monte Carlo: An Adaptive Subsampling Approach. In *International Conference on Machine Learning (ICML). Proceedings of the 31st International Conference on Machine Learning (ICML)* 405–413.

BERNARD-MICHEL, C., DOUTÉ, S., GARDES, L. and GIRARD, S. (2007). Estimation of Mars Surface Physical Properties from Hyperspectral Images Using Sliced Inverse Regression.

BERNARD-MICHEL, C., DOUTÉ, S., FAUVEL, M., GARDES, L. and GIRARD, S. (2009). Retrieval of Mars Surface Physical Properties from OMEGA Hyperspectral Images Using Regularized Sliced Inverse Regression. *Journal of Geophysical Research: Planets* **114** E06005.

BERTRAND, C., OHMI, M., SUZUKI, R. and KADO, H. (2001). A Probabilistic Solution to the MEG Inverse Problem via MCMC Methods: The Reversible Jump and Parallel Tempering Algorithms. *IEEE Transactions on Biomedical Engineering* **48** 533–542.

BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review* **59** 65–98.

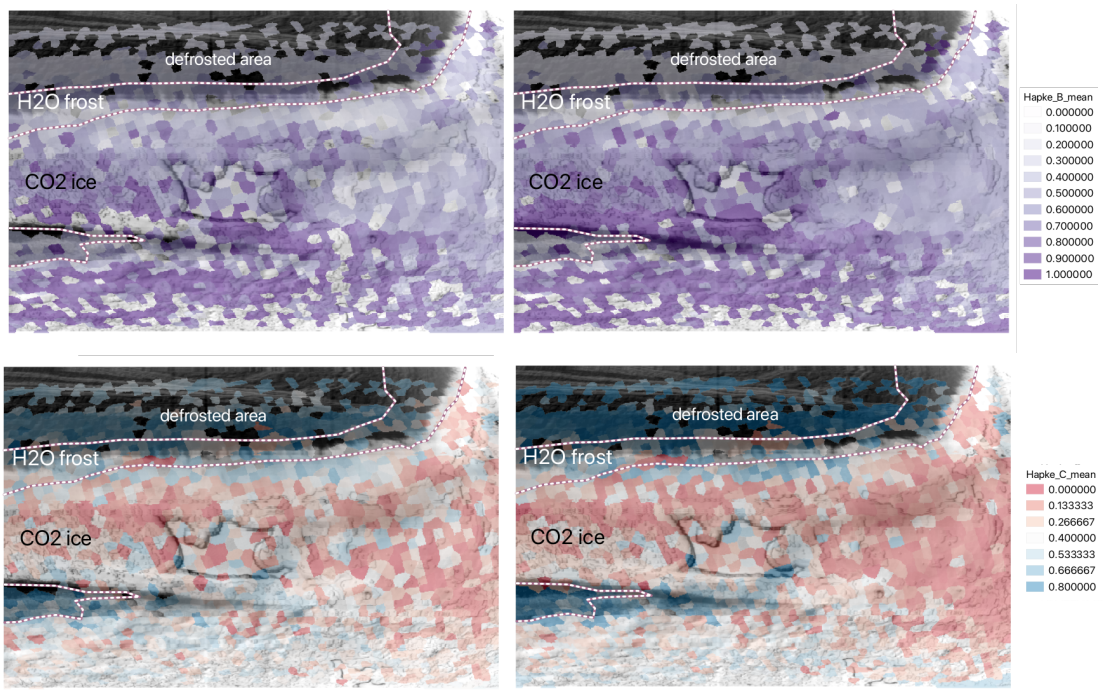

FIG 9. *Mars South pole dataset. Parameters b (top) and c (bottom) averaged over spectral dimension, predicted using* $\bar{\mathbf{x}}_{IS-G}$ *(left) or* $\hat{\mathbf{x}}_{best}$ *(right).*

CARREIRA-PERPINAN, M. A. (2000). Mode-Finding for Mixtures of Gaussian Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** 1318–1323.

CEAMANOS, X., DOUTÉ, S., FERNANDO, J., SCHMIDT, F., PINET, P. and LYAPUSTIN, A. (2013). Surface Reflectance of Mars Observed by CRISM/MRO: 1. Multi-Angle Approach for Retrieval of Surface Reflectance from CRISM Observations (MARS-ReCO). *Journal of Geophysical Research: Planets* **118** 514–533.

CHIANCONE, A., FORBES, F. and GIRARD, S. (2017). Student Sliced Inverse Regression. *Computational Statistics & Data Analysis* **113** 441–456.

COHEN, T. S., GEIGER, M., KOEHLER, J. and WELLING, M. (2018). Spherical CNNs. *arXiv:1801.10130 [cs, stat]*.

COOK, R. D. and FORZANI, L. (2019). Partial Least Squares Prediction in High-Dimensional Regression. *Annals of Statistics* **47** 884–908.

DARVISHZADEH, R., MATKAN, A. A. and AHANGAR, A. D. (2012). Inversion of a radiative transfer model for estimation of rice canopy chlorophyll content using a lookup-table approach. *IEEE Journal of selected topics in applied earth observations and remote sensing* **5** 1222–1230.

DELEFORGE, A., FORBES, F. and HORAUD, R. (2015). High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables. *Statistics and Computing* **25** 893–911.

DELEFORGE, A., FORBES, F., BA, S. and HORAUD, R. (2015). Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies. *IEEE Journal of Selected Topics in Signal Processing* **9** 1037–1048.

Douté, S. and Pilorget, C. (2017). Physical State and Temporal Evolution of Icy Surfaces in the Mars South Pole. *European Planetary Science Congress* **11** EPSC2017-491.

Fernando, J., Schmidt, F. and Douté, S. (2016). Martian Surface Microtexture from Orbital CRISM Multi-Angular Observations: A New Perspective for the Characterization of the Geological Processes. *Planetary and Space Science* **128** 30–51.

Frau-Pascual, A., Vincent, T., Sloboda, J., Ciuciu, P. and Forbes, F. (2014). Physiologically Informed Bayesian Analysis of ASL fMRI Data. In *Bayesian and grAphical Models for Biomedical Imaging* (M. J. Cardoso, I. Simpson, T. Arbel, D. Precup and A. Ribbens, eds.). *Lecture Notes in Computer Science* 37–48. Springer International Publishing.

Giovannelli, J.-F. and Idier, J. (2015). *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*. John Wiley & Sons, Inc.

Golbabaee, M., Chen, D., Gómez, P. A., Menzel, M. I. and Davies, M. (2019). Geometry of Deep Learning for Magnetic Resonance Fingerprinting. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* **2019** 7825–7829.

Hennig, C. (2010). Methods for Merging Gaussian Mixture Components. *Advances in Data Analysis and Classification* **4** 3–34.

Hoppe, E., Körzdörfer, G., Würfl, T., Wetzl, J., Lugauer, F., Pfeuffer, J. and Maier, A. (2017). Deep Learning for Magnetic Resonance Fingerprinting: A New Approach for Predicting Quantitative Parameter Values from Time Series. *Studies in Health Technology and Informatics* **243** 202–206.

Hovorka, R., Canonico, V., Chassin, L. J., Haueter, U., Massi-Benedetti, M., Federici, M. O., Pieber, T. R., Schaller, H. C., Schaupp, L., Vering, T. and Wilinska, M. E. (2004). Nonlinear Model Predictive Control of Glucose Concentration in Subjects with Type 1 Diabetes. *Physiological Measurement* **25** 905–920.

Ingrassia, S., Minotti, S. C. and Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of classification* **29** 363–401.

Izbicki, R., Lee, A. B. and Pospisil, T. (2019). ABC–CDE: Toward Approximate Bayesian Computation with Complex High-Dimensional Data and Limited Simulations. *Journal of Computational and Graphical Statistics* **28** 481–492.

Lathuiliere, S., Juge, R., Mesejo, P., Munoz-Salinas, R. and Horaud, R. (2017). Deep Mixture of Linear Inverse Regressions Applied to Head-Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4817–4825.

Lemasson, B., Pannetier, N., Coquery, N., Boisserand, L. S. B., Collomb, N., Schuff, N., Moseley, M., Zaharchuk, G., Barbier, E. L. and Christen, T. (2016). MR Vascular Fingerprinting in Stroke and Brain Tumors Models. *Scientific Reports* **6** 37071.

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* **86** 316–327.

Ma, D., Gulani, V., Seiberlich, N., Liu, K., Sunshine, J. L., Duerk, J. L. and Griswold, M. A. (2013). Magnetic Resonance Fingerprinting. *Nature* **495** 187–192.

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

Mesejo, P., Saillet, S., David, O., Bénar, C., Warnking, J. M. and Forbes, F. (2016). A Differential Evolution-Based Approach for Fitting a Nonlinear Biophysical Model to fMRI BOLD Data. *IEEE Journal of Selected Topics in Signal Processing* **10** 416–427.

Murchie, S. L., Seelos, F. P., Hash, C. D., Humm, D. C., Malaret, E., McGovern, J. A.,

Choo, T. H., Seelos, K. D., Buczkowski, D. L., Morgan, M. F., Barnouin-Jha, O. S., Nair, H., Taylor, H. W., Patterson, G. W., Harvel, C. A., Mustard, J. F., Arvidson, R. E., McGuire, P., Smith, M. D., Wolff, M. J., Titus, T. N., Bibring, J.-P. and Poulet, F. (2009). Compact Reconnaissance Imaging Spectrometer for Mars Investigation and Data Set from the Mars Reconnaissance Orbiter's Primary Science Phase. *Journal of Geophysical Research: Planets* **114** E00D07.

Nataraj, G., Nielsen, J.-F., Scott, C. and Fessler, J. A. (2018). Dictionary-Free MRI PERK: Parameter Estimation via Regression with Kernels. *IEEE Transactions on Medical Imaging* **37** 2103–2114.

Nguyen, H. D., Chamroukhi, F. and Forbes, F. (2019). Approximation Results Regarding the Multiple-Output Gaussian Gated Mixture of Linear Experts Model. *Neurocomputing* **366** 208–214.

Perthame, E., Forbes, F. and Deleforge, A. (2018). Inverse Regression Approach to Robust Nonlinear High-to-Low Dimensional Mapping. *Journal of Multivariate Analysis* **163** 1–14.

Pilorget, C., Fernando, J., Ehlmann, B. L., Schmidt, F. and Hiroi, T. (2016). Wavelength Dependence of Scattering Properties in the VIS–NIR and Links with Grain-Scale Physical and Compositional Properties. *Icarus* **267** 296–314.

Potin, S., Beck, P., Schmitt, B. and Moynier, F. (2019). Some Things Special about NEAs: Geometric and Environmental Effects on the Optical Signatures of Hydration. *Icarus* **333** 415–428.

Ray, S. and Lindsay, B. G. (2005). The Topography of Multivariate Normal Mixtures. *The Annals of Statistics* **33** 2042–2065.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York.

Runnalls, A. R. (2007). Kullback-Leibler Approach to Gaussian Mixture Reduction. *Aerospace and Electronic Systems, IEEE Transactions on* **43** 989–999.

Schmidt, F. and Fernando, J. (2015). Realistic Uncertainties on Hapke Model Parameters from Photometric Measurement. *Icarus* **260** 73–93.

Sisson, S. A., Fan, Y. and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. CRC Press.

Sobol, I. M. (1967). On the Distribution of Points in a Cube and the Approximate Evaluation of Integrals. *USSR Computational Mathematics and Mathematical Physics* **7** 86–112.

Song, P., Eldar, Y. C., Mazor, G. and Rodrigues, M. R. D. (2019). HYDRA: Hybrid Deep Magnetic Resonance Fingerprinting. *Medical Physics* **46** 4951–4969.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014). Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. *Other Titles in Applied Mathematics*. Society for Industrial and Applied Mathematics.

Tarantola, A., Valette, B. et al. (1982). Inverse Problems= Quest for Information. *Journal of Geophysics— IF 32.18* **50** 159–170.

Tu, C.-C., Forbes, F., Lemasson, B. and Wang, N. (2019). Prediction with High Dimensional Regression via Hierarchically Structured Gaussian Mixtures and Latent Variables. *Journal of the Royal Statistical Society: Series C Applied Statistics* **68** 1485–1507.

Virtue, P., Yu, S. X. and Lustig, M. (2017). Better than Real: Complex-Valued Neural Nets for MRI Fingerprinting. In *2017 IEEE International Conference on Image Processing (ICIP)*

3953–3957.

Zhao, B., Setsompop, K., Ye, H., Cauley, S. F. and Wald, L. L. (2016). Maximum Likelihood Reconstruction for Magnetic Resonance Fingerprinting. *IEEE transactions on medical imaging* **35** 1812–1823.