VARIATIONAL BAYES' METHOD FOR FUNCTIONS WITH APPLICATIONS TO SOME INVERSE PROBLEMS

JUNXIONG JIA*, QIAN ZHAO†, ZONGBEN XU‡, DEYU MENG§, AND YEE LEUNG¶

Abstract. Bayesian approach, a useful tool for quantifying uncertainties, has been extensively employed to solve the inverse problems of partial differential equations (PDEs). One of the main difficulties in employing the Bayesian approach to such problems is how to extract information from the posterior probability measure. Compared with conventional sampling-type methods, variational Bayes' method (VBM) has been intensively examined in the field of machine learning attributed to its ability in extracting approximately the posterior information with lower computational cost. In this paper, we generalize the conventional finite-dimensional VBM to the infinite-dimensional space rigorously solve the inverse problems of PDEs. We further establish a general infinite-dimensional mean-field approximate theory and apply it to the linear inverse problems under the Gaussian and Laplace noise assumptions at the abstract level. The results of some numerical experiments substantiate the effectiveness of the proposed approach.

Key words. inverse problems, variational Bayes' method, mean-field approximation, machine learning, inverse source problem

AMS subject classifications. 65L09, 49N45, 62F15

1. Introduction.

Motivated by the significant applications in medical imaging, seismic explorations and many other domains, the field of inverse problems has undergone an enormous development over the past few decades. In handling an inverse problem, we usually meet ill-posed issue in the sense that the solution lacks stability or even uniqueness [31, 49]. The regularization approach, including Tikhonov and Total-Variation regularization, is one of the most popular approaches to alleviate this ill-posed issue of inverse problems. In the regularization approach, statistical models for data are mostly employed to justify the choice of data discrepancy and for selecting an appropriate regularization parameter. In addition, the statistical properties of data can be investigated carefully, which can be useful for uncertainty quantification [9]. However, statistical assumptions on the model parameters are rarely considered in functional analytic regularization. For a complete review, we refer to Sections 2 and 3 in [4].

The Bayesian inverse approach provides a flexible framework that solves inverse problems by transforming them into statistical inference problems, thereby making it feasible to analyze the uncertainty of the solutions to the inverse problems. Inverse problems are usually accompanied by a forward operator originating from some partial differential equations (PDEs), thereby introducing difficulties to the direct use of the finite-dimensional Bayes' formula. The following two strategies can be employed to solve this problem:

1. Discretize-then-Bayesianize: The PDEs are initially discretized to approximate the original problem in some finite-dimensional space, and the reduced

*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China; (jjx323@xjtu.edu.cn).

†School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China; (timmy.zhaoqian@xjtu.edu.cn).

‡School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China; (zbxu@mail.xjtu.edu.cn).

§School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China; (dymeng@mail.xjtu.edu.cn).

¶ The Chinese University of Hong Kong, Shatin, Hong Kong, SAR (yeeleung@cuhk.edu.hk).

approximate problem is then solved by using the Bayes' method.

2. Bayesianize-then-discretize: The Bayes' formula and algorithms are initially constructed on infinite-dimensional space, and after the infinite-dimensional algorithm is built, some finite-dimensional approximation is carried out.

The first strategy makes available all the Bayesian inference methods developed in the statistical literature [35]. However, given that the original problems are defined on infinite-dimensional space, several problems, such as non-convergence and dimensional dependence, tend to emerge when using this strategy [16, 36]. By employing the second strategy, the discretization-invariant property naturally holds given that the Bayes' formula and algorithms are properly defined on some separable Banach space [19, 48]. In the following sections, we confine ourselves to the second strategy, that is, postponing the discretization to the final step.

One of the essential issues for employing the Bayes' inverse method is how to extract information from the posterior probability measure. Previous studies have adopted two major approaches to address such issue, namely, the point estimate method and the sampling method. For the point estimate method, the maximum a posteriori (MAP) estimate, which is intuitively equivalent to solving an optimization problem, is often utilized. The intuitive equivalence relation has been rigorously analyzed recently [2, 13, 18, 21, 27]. In some situations [32, 49], MAP estimates are more desirable and computationally feasible than the entire posterior distribution. However, point estimates cannot provide uncertainty quantification and are usually recognized as incomplete Bayes' method.

To extract all information encoded in the posterior distribution, sampling methods, such as the Markov chain Monte Carlo (MCMC), are often employed. In 2013, Cotter et al. [16] proposed using the MCMC method for functions to ensure that the convergence speed of the algorithm is robust under mesh refinement. Multiple dimension-independent MCMC-type algorithms have also been proposed [17, 23]. Although MCMC is highly-efficient as a sampling method, its computational cost is unacceptable for many applications, including the full waveform inversion [24].

In this paper, we aim to propose a variational method that can perform uncertainty analysis at a computational cost which is comparable to that for computing the MAP estimates. For finite-dimensional problems, such types of methods, named as variational Bayes' methods (VBM), have been broadly investigated in the field of machine learning [8, 41, 52, 53]. In addressing the inverse problems, Jin et al. [34, 33] employed VBM to investigate a hierarchical formulation of the finite-dimensional inverse problems when the noise is distributed according to Gaussian or centered-t distribution. Guhua et al. [26] generalized this method further to the case when the noise is distributed according to skewed-t error distribution. Finite-dimensional VBM has been recently applied to study the porous media flows in heterogeneous stochastic media [51].

All the aforementioned investigations are conducted based on finite-dimensional VBM. Therefore, only the first strategy as aforementioned can be employed to solve the inverse problems. To the best of our knowledge, only two relevant works have investigated VBM under the infinite-dimensional setting. Specifically, when the approximate probability measures are restricted to be Gaussian, Pinski et al. [44, 45] employed a calculus-of-variations viewpoint to study the properties of Gaussian approximate sequences with Kullback-Leibler (KL) divergence as the a fitness measure. Relying on the Robbins-Monro algorithm, they developed a novel algorithm for obtaining the approximate Gaussian measure. Until now, no study has been conducted beyond such Gaussian approximate measure assumption. However, various approxi-

mate probability measures have been frequently used for training deep neural networks and solving finite-dimensional inverse problems [34, 33]. In this case, for applications in inverse problems concerned with PDEs, a VBM with approximate measures other than Gaussian should be necessarily constructed on infinite-dimensional space.

In the following, we focus on the classical mean-field approximation that is widely employed for the finite-dimensional case. This approximation originally stems from the theory of statistical mechanics for treating many-body systems. Inspired by finite-dimensional theory, we construct a general infinite-dimensional mean-field approximate based VBM, which allows the use of general approximate probability measures beyond Gaussian. Examples are also given to illustrate the flexibility of our proposed approach. The contributions of our work can be summarized as follows:

- By introducing a reference probability measure and using the calculus of variations, we establish a general mean-field approximate based VBM on Hilbert spaces that provides a flexible framework for introducing techniques developed on finite-dimensional space to infinite-dimensional space.
- We apply the proposed theory to a general linear inverse problem (the forward map is assumed to be a bounded linear operator) with Gaussian and Laplace noise assumptions. Precise assumptions can be found in Subsection 3.1. Through detailed calculations, we construct iterative algorithms for functions. To the best of our knowledge, VBM with Laplace noise assumption has not been previously employed for solving inverse problems, even those that are restricted to finite-dimensional space.
- We solve the inverse source problems of Helmholtz equations with multi-frequency data by using the proposed VBM with Gaussian and Laplace noise assumptions. The algorithms not only provide a point estimate but also give the standard deviations of the numerical solutions.

The outline of this paper is as follows. In Section 2, we construct the general infinite-dimensional VBM based on the mean-field approximate assumption. In Section 3, under the hierarchical formulation, we apply the proposed theory to an abstract linear inverse problem with Gaussian and Laplace noise assumptions. In Section 4, we present concrete numerical examples to illustrate the effectiveness of our proposed approach. In Section 5, we summarize our findings and propose some directions for further research. Due to the limited space, we did not provide all proofs in the main text. *All of the proofs are given in the supplemental materials.*

2. General theory on infinite-dimensional space.

In Subsection 2.1, we provide the necessary background of our theory and prove some basic results concerning with the existence of minimizers for finite product probability measures. In Subsection 2.2, we present our infinite-dimensional variational Bayes' approach.

2.1. Existence theory.

In this subsection, we first recall some general facts about the Kullback-Leibler (KL) approximation from the viewpoint of calculus of variations, and then provide some new theorems for product of probability measures that form the basis of our investigation. Let $\mathcal{H}$ be a separable Hilbert space endowed with its Borel sigma algebra $\mathcal{B}(\mathcal{H})$, and let $\mathcal{M}(\mathcal{H})$ be the set of Borel probability measures on $\mathcal{H}$.

For inverse problems, we usually need to find a probability measure μ on $\mathcal{H}$, which is called the posterior probability measure, specified by its density with respect to a prior probability measure μ_0 [48]. Let the Bayesian formula on the Hilbert space be

defined by

$$\frac{d\mu}{d\mu_0}(x) = \frac{1}{Z_\mu}\exp\big(-\Phi(x)\big), \tag{1}$$

where $\Phi(x) : \mathcal{H} \to \mathbb{R}$ is a continuous function, and $\exp\big(-\Phi(x)\big)$ is integrable with respect to μ_0. The constant Z_μ is chosen to ensure that μ is indeed a probability measure.

Let $\mathcal{A} \subset \mathcal{M}(\mathcal{H})$ be a set of "simpler" measures that can be efficiently calculated. Our aim is to find the closest element ν to μ with respect to the KL divergence from subset $\mathcal{A}$. For any $\nu \in \mathcal{M}(\mathcal{H})$ that is absolutely continuous with respect to μ, the KL divergence is defined as

$$D_{\mathrm{KL}}(\nu||\mu) = \int_{\mathcal{H}} \log\left(\frac{d\nu}{d\mu}(x)\right)\frac{d\nu}{d\mu}(x)\mu(dx) = \mathbb{E}^{\mu}\left[\log\left(\frac{d\nu}{d\mu}(x)\right)\frac{d\nu}{d\mu}(x)\right], \tag{2}$$

where the convention $0\log 0 = 0$ has been used. If ν is not absolutely continuous with respect to μ, then the KL divergence is defined as $+\infty$. With this definition, this paper examines the following minimization problem:

$$\arg\min_{\nu\in\mathcal{A}} D_{\mathrm{KL}}(\nu||\mu). \tag{3}$$

There are some studies of the above general minimization problem (3) taken from the perspective of the calculus of variations. We follow this line of investigations in this section, and for the convenience of the readers, we present the following proposition, which has been proven in [45].

PROPOSITION 1. *Let $\mathcal{A}$ be closed with respect to weak convergence. Then, given $\mu \in \mathcal{M}(\mathcal{H})$, assume that there exists $\nu \in \mathcal{A}$ such that $D_{KL}(\nu||\mu) < \infty$. It follows that there exists a minimizer $\nu \in \mathcal{A}$ solving*

$$\arg\min_{\nu\in\mathcal{A}} D_{KL}(\nu||\mu).$$

As stated in the Introduction, we aim to construct a mean-field approximation that usually takes the following factorized form for the finite-dimensional case

$$q(x_1, \cdots, x_M) = \prod_{j=1}^{M} q(x_j), \tag{4}$$

where $q(x_1, \cdots, x_M)$ is the full probability density function, $q(x_j)$ is the probability density function for x_j, and $x_j \in \mathbb{R}^{N_j}(N_j \in \mathbb{N})$ for $j = 1, 2, \cdots, M$. That is, we assume that $x_1, \cdots, x_M$ are independent random variables. By carefully choosing the random variables $\{x_j\}_{j=1}^{M}$, this independence assumption will lead to computationally efficient solutions when conjugate prior probabilities are employed. Additional details can be found in Chapter 9 of [8] and in some recently published papers [33, 52, 53].

Inspired by formula (4), for a fixed positive constant M, we specify the Hilbert space $\mathcal{H}$ and subset $\mathcal{A}$ as

$$\mathcal{H} = \prod_{j=1}^{M} \mathcal{H}_j, \qquad \mathcal{A} = \prod_{j=1}^{M} \mathcal{A}_j, \tag{5}$$

where $\mathcal{H}_j (j = 1, \cdots, M)$ are a series of separable Hilbert space and $\mathcal{A}_j \subset \mathcal{M}(\mathcal{H}_j)$. Let $\nu := \prod_{i=1}^M \nu^i$ be a probability measure such that $\nu(dx) = \prod_{i=1}^M \nu^i(dx)$. With these assumptions, the minimization problem in (3) can be rewritten as

$$\arg\min_{\nu^i \in \mathcal{A}_i} D_{\mathrm{KL}}\left(\prod_{i=1}^M \nu^i || \mu \right) \tag{6}$$

for suitable sets $\mathcal{A}_i$ with $i = 1, 2, \cdots, M$. The general result shown in Proposition 1 indicates that the closedness of the subset $\mathcal{A}$ under weak convergence is crucial for the existence of the approximate measure ν. Therefore, we present the following lemma that illustrates the closedness of $\mathcal{A}$ as defined in (5).

LEMMA 2. *For $i = 1, 2, \cdots, M$, let $\mathcal{A}_i \subset \mathcal{M}(\mathcal{H}_i)$ be a series of sets closed under weak convergence of probability measures. Define*

$$\mathcal{C} := \left\{ \nu := \prod_{j=1}^M \nu^j \;\middle|\; \nu^j \in \mathcal{A}_j \quad \text{for } j = 1, 2, \cdots, M \right\}. \tag{7}$$

Then, the set $\mathcal{C}$ is closed under the weak convergence of probability measures.

From Lemma 2 and Proposition 1, we can prove the following existence result.

THEOREM 3. *For $i = 1, 2, \cdots, M$, let $\mathcal{A}_i$ be closed with respect to weak convergence. Given $\mu \in \mathcal{M}(\prod_{i=1}^M \mathcal{H}_i)$, we assume that there exists $\nu^i \in \mathcal{A}_i$ for $i = 1, \cdots, M$ such that $D_{KL}(\prod_{i=1}^M \nu^i || \mu) < \infty$. Then, there exists a minimizer $\prod_{i=1}^M \nu^i$ that solves problem (6).*

REMARK 4. *In Theorem 3, we only illustrate the existence of the approximate measure ν without uniqueness. When the approximate measures are assumed to be Gaussian, uniqueness has been obtained with the λ-convex requirement of the potential Φ appearing in the Bayes' formula (1) [45]. We cannot expect uniqueness generally even for most of the practical problems defined on the finite-dimensional space. Therefore, we will not pursue the uniqueness results here.*

The result shown in Theorem 3 does not tell us much about the manner in which minimizing sequences approach the limit. After further deductions, we can precisely characterize the convergence.

THEOREM 5. *Let $\{\nu_n = \prod_{j=1}^M \nu_n^j\}_{n=1}^\infty$ be a sequence in $\prod_{j=1}^M \mathcal{M}(\mathcal{H}_j)$, and let $\nu_* = \prod_{j=1}^M \nu_*^j \in \prod_{j=1}^M \mathcal{M}(\mathcal{H}_j)$ and $\mu \in \mathcal{M}(\prod_{j=1}^M \mathcal{H}_j)$ be probability measures such that for any $n \geq 1$, we have $D_{KL}(\nu_n || \mu) < \infty$ and $D_{KL}(\nu_* || \mu) < \infty$. Suppose that ν_n converges weakly to ν_* and $\nu_n^j \ll \nu_*^j$ for $j = 1, 2, \cdots, M$ and that*

$$D_{KL}(\nu_n || \mu) \to D_{KL}(\nu_* || \mu). \tag{8}$$

Then, ν_n^j converges to ν_*^j in the total variation norm for $j = 1, 2, \cdots, M$.

Combining Theorems 3 and 5, we immediately obtain the following result.

COROLLARY 6. *For $j = 1, 2, \cdots, M$, let $\mathcal{A}_j \subset \mathcal{M}(\mathcal{H}_j)$ be closed with respect to weak convergence. Given $\mu \in \mathcal{M}(\prod_{j=1}^M \mathcal{H}_j)$, there exists $\nu = \prod_{j=1}^M \nu^j \in \prod_{j=1}^M \mathcal{A}_j$ with $D_{KL}(\nu || \mu) < \infty$. Let $\nu_n = \prod_{j=1}^M \nu_n^j \in \prod_{j=1}^M \mathcal{A}_j$ satisfy*

$$D_{KL}(\nu_n || \mu) \to \inf_{\nu \in \prod_{j=1}^M \mathcal{A}_j} D_{KL}(\nu || \mu). \tag{9}$$

Then, after passing to a subsequence, we have

- *ν_n converges weakly to $\nu_* = \prod_{j=1}^M \nu_*^j \in \prod_{j=1}^M \mathcal{M}(\mathcal{H}_j)$ that realizes the infimum in (9);*
- *each ν_n^j converges weakly to ν_*^j for $j = 1, 2, \cdots, M$.*

In addition, for $j = 1, 2, \cdots, M$, if $\nu_n^j \ll \nu_*^j$ for all n, each ν_n^j converges to ν_*^j in the total-variation norm.

REMARK 7. *Because our results rely on conclusions given in [45] that hold on Polish spaces, it should be pointed out that all of the theoretical results presented in this subsection actually hold on Polish spaces.*

2.2. Mean-field approximation for functions. For finite-dimensional cases, the mean-field approximation has been widely employed in the field of machine learning. On the basis of the results presented in Subsection 2.1, we construct a mean-field approximation approach on infinite-dimensional space, which will be useful for solving the inverse problems of PDEs.

In the previous work, e.g., Examples 3.8 and 3.9 in [45] and the general setting described in [44], their idea is replacing the classical density functions by the density functions with respect to the prior measure. In [44, 45], prior measures are taken to be Gaussian measures, which take the role played by the Lebesgue measure in the finite-dimensional setting, as a reference measure. Inspired by these studies, we may assume that the approximate probability measure ν introduced in (3) is equivalent to μ_0 defined by

$$\frac{d\nu}{d\mu_0}(x) = \frac{1}{Z_\nu} \exp\big(-\Phi_\nu(x)\big). \tag{10}$$

Compared with the finite-dimensional case, a natural way for introducing an independence assumption is to assume that the potential $\Phi_\nu(x)$ can be decomposed as

$$\exp\left(-\Phi_\nu(x)\right) = \prod_{j=1}^M \exp\left(-\Phi_\nu^j(x_j)\right), \tag{11}$$

where $x = (x_1, \cdots, x_M)$. However, this intuitive idea prevents us from incorporating those parameters contained in the prior probability measure into the hierarchical Bayes' model that is used in finite-dimensional cases [33, 53]. Given these considerations, we propose the following assumption that introduces a reference probability measure.

ASSUMPTIONS 8. *Let us introduce a reference probability measure*

$$\mu_r(dx) = \prod_{j=1}^M \mu_r^j(dx_j), \tag{12}$$

which is equivalent to the prior probability measure with the following relation being true:

$$\frac{d\mu_0}{d\mu_r}(x) = \frac{1}{Z_0} \exp(-\Phi^0(x)). \tag{13}$$

For each $j = 1, 2, \cdots, M$, there is a predefined continuous function $a_j(\epsilon, x_j)$[1] where ϵ is a positive number and $x_j \in \mathcal{H}_j$. Concerning these functions, we assume that

[1]These functions naturally appear when considering concrete examples, which will be specified in Remark 15. In the last part of the supplementary materials, we provide a detailed illustration of the Gaussian noise example, which may provide more intuitions.

$\mathbb{E}^{\mu_r^j}[a_j(\epsilon,\cdot)] < \infty$ *where* $\epsilon \in [0, \epsilon_0^j)$ *with* ϵ_0^j *is a small positive number* $(j = 1, \cdots, M)$. *We also assume that the approximate probability measure* ν *is equivalent to the reference measure* μ_r *and that the Radon-Nikodym derivative of* ν *with respect to* μ_r *takes the following form*

$$\frac{d\nu}{d\mu_r}(x) = \frac{1}{Z_r} \exp\Big(- \sum_{j=1}^{M} \Phi_j^r(x_j)\Big). \tag{14}$$

Following Assumptions 8, we know that the approximate measure can be decomposed as $\nu(dx) = \prod_{j=1}^{M} \nu^j(dx_j)$ with

$$\frac{d\nu^j}{d\mu_r^j} = \frac{1}{Z_r^j} \exp\big(- \Phi_j^r(x_j)\big). \tag{15}$$

Here, $Z_r^j = \mathbb{E}^{\mu_r^j}\big(\exp\big(- \Phi_j^r(x_j)\big)\big)$ ensures that ν^j is indeed a probability measure.

REMARK 9. *The reference measure introduced above can be easily specified for concrete examples. Fix a component* j, *if* x_j *belongs to some finite-dimensional Hilbert space, we assume that the prior measure of* x_j *has a density function* $p(\cdot)$. *Then we can choose the reference measure of* x_j *just equal to the prior measure. Formula (15) for this component reduces to the classical finite-dimensional case. If* x_j *belongs to some Hilbert space with the prior measure contains some hyper-parameters, there may be no universal strategies for choosing the reference measure. Here, we provide a simple example to give some intuitions. Assume* $x_j \sim \mathcal{N}(0, \mathcal{C}_\tau)$ *with* $\mathcal{C}_\tau := (\tau^2 I - \Delta)^{-\alpha}$ *(*α *is a fixed positive number) [20], we can choose the reference measure to be a Gaussian measure* $\mathcal{N}(0, \mathcal{C})$ *with* $\mathcal{C} := (I - \Delta)^{-\alpha}$, *which is equivalent to* $\mathcal{N}(0, \mathcal{C}_\tau)$ *under some appropriate conditions (rigorous results are given in Theorem 1 in [20]).*

For convenience, let us introduce some notations. For $j = 1, 2, \cdots, M$, let $\mathcal{Z}_j$ be defined as a Hilbert space that is embedded in $\mathcal{H}_j$. Denote C_N be a positive constant related to N. Then, for $j = 1, 2, \cdots, M$, we introduce

$$\mathrm{R}_j^1 = \Big\{ \Phi_j^r \,\Big|\, \sup_{1/N \le \|x_j\|_{\mathcal{Z}_j} \le N} \Phi_j^r(x_j) \le C_N < \infty \text{ for all } N > 0 \Big\},$$

$$\mathrm{R}_j^2 = \Big\{ \Phi_j^r \,\Big|\, \int_{\mathcal{H}_j} \exp\big(-\Phi_j^r(x_j)\big) \max(1, a_j(\epsilon, x_j)) \mu_r^j(dx_j) \le C < \infty, \text{ for } \epsilon \in [0, \epsilon_0^j) \Big\},$$

where C is an arbitrary large positive constant, ϵ_0^j and $a_j(\cdot,\cdot)$ are defined as in Assumptions 8. With these preparations, we can define $\mathcal{A}_j$ $(j = 1, 2, \cdots, M)$ as follows:

$$\mathcal{A}_j = \left\{ \nu^j \in \mathcal{M}(\mathcal{H}_j) \,\middle|\, \begin{array}{l} \nu^j \text{ is equivalent to } \mu_r^j \text{ with (15) holds true,} \\ \text{and } \Phi_j^r \in \mathrm{R}_j^1 \cap \mathrm{R}_j^2 \end{array} \right\}. \tag{16}$$

Before using Theorem 3, we need to illustrate the closedness of $\mathcal{A}_j$ $(j = 1, 2, \cdots, M)$ under the weak convergence topology. Actually, we can prove the desired results shown blow.

THEOREM 10. *For* $j = 1, 2, \cdots, M$, *we denote* $T_N^j = \{x_j \,|\, 1/N \le \|x_j\|_{\mathcal{Z}_j} \le N\}$, *with* N *being an arbitrary positive constant. For each reference measure* μ_r^j, *we assume that* $\sup_N \mu_r^j(T_N^j) = 1$. *Then, each* $\mathcal{A}_j$ *is closed with respect to weak convergence and problem (6) possesses a solution* $\prod_{j=1}^{M} \nu^j$ *with* $\nu^j \in \mathcal{A}_j$ *for* $j = 1, 2, \cdots, M$.

In the following theorem, we provide a special form of solution that helps us obtain the optimal approximate measure via simple iterative updates.

THEOREM 11. *Assume that the approximate probability measure in problem (6) satisfies Assumptions 8 and the assumptions presented in Theorem 10. Using the same notations as in Theorem 10, in addition, we assume*

$$\sup_{x_i\in T_N^i}\sup_{\substack{\nu^j\in\mathcal{A}_j\\ j\neq i}}\int_{\prod_{j\neq i}\mathcal{H}_j}\big(\Phi^0(x)+\Phi(x)\big)\,1_A(x)\prod_{j\neq i}\nu^j(dx_j)<\infty, \tag{17}$$

and

$$\sup_{\substack{\nu^j\in\mathcal{A}_j\\ j\neq i}}\int_{\mathcal{H}_i}\exp\bigg(-\int_{\prod_{j\neq i}\mathcal{H}_j}(\Phi^0(x)+\Phi(x))1_{A^c}(x)\prod_{j\neq i}\nu^j(dx_j)\bigg)M_i(x)\mu_r^i(dx_i)<\infty, \tag{18}$$

where $A:=\{x\,|\,\Phi^0(x)+\Phi(x)\geq 0\}$, *and* $M_i(x):=\max\left(1,a_i(\epsilon,x_i)\right)$ *with* $i,j=1,2,\cdots,M$. *Then, problem (6) possesses a solution* $\nu=\prod_{j=1}^M\nu^j\in\mathcal{M}(\mathcal{H})$ *with the following form*

$$\frac{d\nu}{d\mu_r}\propto\exp\bigg(-\sum_{i=1}^M\Phi_i^r(x_i)\bigg), \tag{19}$$

where

$$\Phi_i^r(x_i)=\int_{\prod_{j\neq i}\mathcal{H}_j}\bigg(\Phi^0(x)+\Phi(x)\bigg)\prod_{j\neq i}\nu^j(dx_j)+Const \tag{20}$$

and

$$\nu^i(dx_i)\propto\exp\big(-\Phi_i^r(x_i)\big)\mu_r^i(dx_i). \tag{21}$$

REMARK 12. *For* $i=1,2,\cdots,M$, *conditions (17) and (18) ensure that each components of the approximate measure* ν *and the reference probability measure* μ_r *are equivalent. These two conditions can be verified in a straightforward manner for specific examples relying on the integrability and boundedness conditions of* Φ_i^r *contained in the definition of* $\mathcal{A}_i$ *in (16) for* $i=1,2,\cdots,M$.

REMARK 13. *Formula (20) means that the logarithm of the optimal solution for factor* ν^j *can be obtained simply by considering the logarithm of the joint distribution over all of the other variables and then taking the expectation with respect to all of the other factors* $\{\nu^i\}$ *fixed for* $i\neq j$. *This result is in accordance with the finite-dimensional case illustrated in Subsection 2.3 of [52].*

REMARK 14. *Based on Theorem 11, we can therefore seek a solution by first initializing all of the potentials* Φ_j^r *appropriately and then cycling through the potentials and replacing each in turn with a revised estimate given by the right-hand side of (20) evaluated by using the current estimates for all of the other potentials.*

3. Applications to some general inverse problems. In Subsection 3.1, we apply our general theory to an abstract linear inverse problem (ALIP). We assume that the prior and noise probability measures are all Gaussian with some hyper-parameters, and then we formulate hierarchical models that can be efficiently solved by using the variational Bayes' approach. In Subsection 3.2, we assume that the noise is distributed according to the Laplace distribution. Through this assumption, we can formulate algorithms that solve ALIP and are robust to outliers.

3.1. Linear inverse problems with Gaussian noise. In this subsection, we apply our general theory to an abstract linear inverse problem. A detailed investigation of the corresponding finite-dimensional case can be found in [34].

Let $\mathcal{H}_u$ be some separable Hilbert space and N_d be a positive integer. We describe the linear inverse problem as

$$d = Hu + \epsilon, \tag{22}$$

where $d \in \mathbb{R}^{N_d}$ is the measurement data, $u \in \mathcal{H}_u$ is the sought-for solution, H is a bounded linear operator from $\mathcal{H}_u$ to $\mathbb{R}^{N_d}$, and ϵ is a Gaussian random vector with zero mean and $\tau^{-1}I$ variance. We will focus on the hyper-parameter treatment within hierarchical models and the challenges in efficiently exploring the posterior probability.

To formulate this problem under the Bayesian inverse framework, we introduce a prior probability measure for the unknown function u. Let $\mathcal{C}_0$ be a symmetric, positive definite and trace class operator defined on $\mathcal{H}_u$, and let (e_k, α_k) be an eigen-system of the operator $\mathcal{C}_0$ such that $\mathcal{C}_0 e_k = \alpha_k e_k$. Without loss of generality, we assume that the eigenvectors $\{e_k\}_{k=1}^{\infty}$ are orthonormal and the eigenvalues $\{\alpha_k\}_{k=1}^{\infty}$ are in a descending order. In the following, for a function $u \in \mathcal{H}_u$, we denote $u_j := \langle u, e_j \rangle$ for $j = 1, 2, \cdots$. According to Subsection 2.4 in [19], we have

$$\mathcal{C}_0 = \sum_{j=1}^{\infty} \alpha_j e_j \otimes e_j, \tag{23}$$

where $\otimes$ denotes the tensor product on Hilbert space [37, 47]. As indicated in [16, 17, 23], we assume that the data are only informative on a finite number of directions in $\mathcal{H}_u$. Under this assumption, we introduce a positive integer K, which represents the number of dimensions that is informed by the data (i.e., the so-called intrinsic dimensionality), which is different from the discretization dimensionality, i.e., the number of mesh points used to represent the unknown variables. The value of K can be specified with a heuristic approach [23]:

$$K = \min \left\{ k \in \mathbb{N} \,\middle|\, \frac{\alpha_k}{\alpha_1} < \epsilon \right\}, \tag{24}$$

where ϵ is a prescribed threshold. Let λ be a positive real number, then, we define

$$\mathcal{C}_0^K(\lambda) := \sum_{j=1}^{K} \lambda^{-1} \alpha_j e_j \otimes e_j + \sum_{j=K+1}^{\infty} \alpha_j e_j \otimes e_j, \tag{25}$$

which is obviously a symmetric, positive definite and trace-class operator. Numerical results shown in [23] indicate that the above heuristic approach could provide acceptable results when ϵ is small enough for a lot of practical inverse problems. However, if the data is particularly informative and far from the prior, this heuristic approach may lead to a Bayesian inference model that does not adequately incorporate information encoded in data. Concerned with this problem, we intend to give more detailed discussions in our future work. We refer to some recent studies [1, 14] that may provide some useful ideas. Then, we assume

$$u \sim \mu_0^{u,\lambda} = \mathcal{N}(u_0, \mathcal{C}_0^K(\lambda)). \tag{26}$$

Let $Gamma(\alpha, \beta)$ be the Gamma probability measure defined on $\mathbb{R}^+$ with the probability density function p_G expressed as

$$p_G(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{27}$$

where $\Gamma(\cdot)$ is the usual Gamma function. Then, except for the function u, we assume that the parameters λ and τ involved in the prior and noise probability measures are all random variables satisfying $\lambda \sim \mu_0^\lambda := Gamma(\alpha_0, \beta_0)$ and $\tau \sim \mu_0^\tau := Gamma(\alpha_1, \beta_1)$. With these preparations, we define the prior probability measure employed for this problem as follows:

$$\mu_0(du, d\lambda, d\tau) = \mu_0^{u,\lambda}(du)\mu_0^\lambda(d\lambda)\mu_0^\tau(d\tau). \tag{28}$$

Let μ be the posterior probability measure for random variables u, λ, and τ. According to Theorems 15 and 16 proved in [19], this probability measure can be defined as

$$\frac{d\mu}{d\mu_0}(u, \lambda, \tau) = \frac{1}{Z_\mu} \tau^{N_d/2} \exp\bigg(-\frac{\tau}{2}\|Hu - d\|^2 \bigg), \tag{29}$$

where

$$Z_\mu = \int_{\mathcal{H}_u \times \mathbb{R}^+ \times \mathbb{R}^+} \tau^{N_d/2} \exp\bigg(-\frac{\tau}{2}\|Hu - d\|^2 \bigg) \mu_0(du, d\lambda, d\tau). \tag{30}$$

To apply the general theory developed in Section 2, we specify the following reference probability measure [2]

$$\mu_r(du, d\lambda, d\tau) = \mu_r^u(du)\mu_r^\lambda(d\lambda)\mu_r^\tau(d\tau), \tag{31}$$

where $\mu_r^u = \mathcal{N}(u_0, \mathcal{C}_0)$ is a Gaussian probability measure, and μ_r^λ and μ_r^τ are chosen to be μ_0^λ and μ_0^τ, respectively.

In Assumption 8, we assume that the approximate probability measure is separable with respect to the random variables u, λ, and τ with the form

$$\nu(du, d\lambda, d\tau) = \nu^u(du)\nu^\lambda(d\lambda)\nu^\tau(d\tau). \tag{32}$$

In addition, we assume that its Radon-Nikodym derivative with respect to μ_r can be written as

$$\frac{d\nu}{d\mu_r}(u, \lambda, \tau) = \frac{1}{Z_r} \exp\bigg(-\Phi_u^r(u) - \Phi_\lambda^r(\lambda) - \Phi_\tau^r(\tau) \bigg). \tag{33}$$

For the Radon-Nikodym derivative of μ_0 with respect to μ_r, we have

$$\begin{aligned} \frac{d\mu_0}{d\mu_r}(u, \lambda, \tau) &= \frac{d\mu_0^{u,\lambda}}{d\mu_r^u}(u) \frac{d\mu_0^\lambda}{d\mu_r^\lambda}(\lambda) \frac{d\mu_0^\tau}{d\mu_r^\tau}(\tau) \\ &= \lambda^{K/2} \exp\bigg(-\frac{1}{2}\|(\mathcal{C}_0^K(\lambda))^{-1/2}(u - u_0)\|^2 + \frac{1}{2}\|\mathcal{C}_0^{-1/2}(u - u_0)\|^2 \bigg) \\ &= \lambda^{K/2} \exp\bigg(-\frac{1}{2}\sum_{j=1}^K (u_j - u_{0j})^2 (\lambda - 1)\alpha_j^{-1} \bigg), \end{aligned} \tag{34}$$

[2] In practical machine learning applications, especially for large-scale scenarios, researchers often assume the approximating measures are independent in each component (a fully diagonal approximation to the posterior) that further reduce of computational burden. This, however, also tends to decrease the computation accuracy due to the neglecting of existing correlations between different components of u. We thus preserve such correlation in our method to alleviate the possible negative influence of ignoring such beneficial knowledge.

which implies that Φ^0 introduced in Assumption 8 takes the following form:

$$\Phi^0(u,\lambda,\tau) = \frac{1}{2}\sum_{j=1}^{K}(u_j - u_{0j})^2(\lambda-1)\alpha_j^{-1} - \frac{K}{2}\log\lambda. \tag{35}$$

REMARK 15. *It should be noted that $\mathbb{R}^+$ is not a Hilbert space. However, the general theory is constructed on some separable Hilbert spaces. This issue can be resolved by considering $\lambda' := \log\lambda$ and $\tau' := \log\tau$ instead of λ and τ. Through this simple transformation, the space of hyper-parameters becomes $\mathbb{R}$ which is a Hilbert space. The calculations presented here also hold true when considering λ' and τ' as hyper-parameters. Actually, we can derive that $e^{\lambda'}$ and $e^{\tau'}$ are distributed according to the same Gamma distributions as λ and τ. Choosing $a_u(\epsilon,u), a_{\lambda'}(\epsilon,\lambda')$, and $a_{\tau'}(\epsilon,\tau')$ appropriately, we can verify the conditions proposed in Theorem 11 (critical steps are provided in the supplementary materials). In the following, we still use λ and τ as hyper-parameters. With this a little abusive use of the general theory (can be rigorously verified through the above simple transformation), the reader may see more clearly the connections between the finite- and infinite-dimensional theory.*

We now calculate $\Phi_u^r(u)$, $\Phi_\lambda^r(\lambda)$, and $\Phi_\tau^r(\tau)$ according to the general results as shown in Theorem 11.

Calculate $\Phi_u^r(u)$: A direct application of formula (20) yields

$$\begin{aligned}\Phi_u^r(u) = \int_0^\infty\int_0^\infty \Big(&\frac{1}{2}\sum_{j=1}^{K}(u_j-u_{0j})^2(\lambda-1)\alpha_j^{-1} + \frac{\tau}{2}\|Hu-d\|^2 \\ &- \frac{K}{2}\log\lambda - \frac{N_d}{2}\log\tau\Big)\nu^\tau(d\tau)\nu^\lambda(d\lambda) + \text{Const} \\ =&\frac{1}{2}\tau^*\|Hu-d\|^2 + \frac{1}{2}(\lambda^*-1)\sum_{j=1}^{K}\alpha_j^{-1}(u_j-u_{0j})^2 + \text{Const},\end{aligned} \tag{36}$$

where

$$\tau^* = \mathbb{E}^{\nu^\tau}[\tau] = \int_0^\infty \tau\nu^\tau(d\tau) \quad \text{and} \quad \lambda^* = \mathbb{E}^{\nu^\lambda}[\lambda] = \int_0^\infty \lambda\nu^\lambda(d\lambda). \tag{37}$$

On the basis of equality (36), we derive

$$\frac{d\nu^u}{d\mu_r^u}(u) \propto \exp\Big(-\frac{\tau^*}{2}\|Hu-d\|^2 - \frac{\lambda^*-1}{2}\sum_{j=1}^{K}\alpha_j^{-1}(u_j-u_{0j})^2\Big). \tag{38}$$

We define

$$\mathcal{C}_0(\lambda^*) = \sum_{j=1}^{K}(\lambda^*)^{-1}\alpha_j e_j\otimes e_j + \sum_{j=K+1}^{\infty}\alpha_j e_j\otimes e_j. \tag{39}$$

Then, according to Example 6.23 in [48], we know that the probability measure ν^u is a Gaussian measure $\mathcal{N}(u^*,\mathcal{C})$ with

$$\mathcal{C}^{-1} = \tau^*H^*H + \mathcal{C}_0(\lambda^*)^{-1} \quad \text{and} \quad u^* = \mathcal{C}\big(\tau^*H^*d + \mathcal{C}_0(\lambda^*)^{-1}u_0\big). \tag{40}$$

Calculate $\Phi_\lambda^r(\lambda)$ and $\Phi_\tau^r(\tau)$: According to formula (20), we have

$$\begin{aligned}\Phi_\lambda^r(\lambda) =& \int_0^\infty \int_{\mathcal{H}_u} \left(\frac{1}{2}\sum_{j=1}^K (u_j - u_{0j})^2 \alpha_j^{-1}\lambda - \frac{K}{2}\log\lambda\right)\nu^u(du)\nu^\tau(d\tau) + \text{Const} \\ =& \frac{1}{2}\mathbb{E}^{\nu^u}\left(\sum_{j=1}^K (u_j - u_{0j})^2 \alpha_j^{-1}\right)\lambda - \frac{K}{2}\log\lambda + \text{Const},\end{aligned} \tag{41}$$

which implies that

$$\frac{d\nu^\lambda}{d\mu_r^\lambda}(\lambda) \propto \lambda^{K/2}\exp\left(-\frac{1}{2}\mathbb{E}^{\nu^u}\left(\sum_{j=1}^K (u_j - u_{0j})^2 \alpha_j^{-1}\right)\lambda\right). \tag{42}$$

Given that λ is a scalar random variable, we can write the density function as bellow:

$$\rho_G(\lambda; \tilde{\alpha}_0, \tilde{\beta}_0) = \frac{\tilde{\beta}_0^{\tilde{\alpha}_0}}{\Gamma(\tilde{\alpha}_0)}\lambda^{\tilde{\alpha}_0 - 1}\exp(-\tilde{\beta}_0\lambda), \tag{43}$$

where

$$\tilde{\alpha}_0 = \alpha_0 + \frac{K}{2} \quad \text{and} \quad \tilde{\beta}_0 = \beta_0 + \frac{1}{2}\mathbb{E}^{\nu^u}\left(\sum_{j=1}^K (u_j - u_{0j})^2 \alpha_j^{-1}\right). \tag{44}$$

Similar to the above calculations of $\Phi_\lambda^r(\lambda)$, we derive

$$\begin{aligned}\Phi_\tau^r(\tau) =& \int_0^\infty \int_{\mathcal{H}_u} \left(\frac{\tau}{2}\|Hu - d\|^2 - \frac{N_d}{2}\log\tau\right)\nu^u(du)\nu^\lambda(d\lambda) + \text{Const} \\ =& \frac{1}{2}\mathbb{E}^{\nu^u}(\|Hu - d\|^2)\tau - \frac{N_d}{2}\log\tau + \text{Const},\end{aligned} \tag{45}$$

which implies

$$\frac{d\nu^\tau}{d\mu_r^\tau}(\tau) \propto \tau^{\frac{N_d}{2}}\exp\left(-\frac{1}{2}\mathbb{E}^{\nu^u}(\|Hu - d\|^2)\tau\right). \tag{46}$$

Therefore, ν^τ is a Gamma distribution $Gamma(\tilde{\alpha}_1, \tilde{\beta}_1)$ with

$$\tilde{\alpha}_1 = \alpha_1 + \frac{N_d}{2} \quad \text{and} \quad \tilde{\beta}_1 = \beta_1 + \frac{1}{2}\mathbb{E}^{\nu^u}(\|Hu - d\|^2). \tag{47}$$

According to the statements in Remark 14, we provide an iterative algorithm based on formulas (40), (43), (44), and (47) in Algorithm 1. Next, we provide a brief discussion of the computational details and the cost of this algorithm. For small- or medium-scale problems, we may construct the finite-dimensional approximate operators H and H^* explicitly [34]. However, for large-scale problems, it is impossible to build finite-dimensional approximations explicitly. Actually, for running the iterations, we only need to compute the mean estimates u_k and some quantities related to ν_k^u such as $\mathbb{E}^{\nu_k^u}(\|Hu - d\|^2)$. For obtaining mean estimates, we can use a matrix-free conjugate gradient (CG) method [10, 25, 43] to solve the following problem

$$(\tau_k H^* H + \mathcal{C}_0(\lambda_k)^{-1})u_k = \tau_k H^* d + \mathcal{C}_0(\lambda_k)^{-1}u_0, \tag{48}$$

Algorithm 1 Variational approximation for the case of Gaussian noise

1: Give an initial guess $\mu_0^{u,\lambda}$ (u_0 and λ), μ_0^{λ} (α_0 and β_0) and μ_0^{τ} (α_1 and β_1). Specify the tolerance *tol* and set $k = 1$.
2: **repeat**
3: Set $k = k + 1$
3: Calculate $\lambda_k = \mathbb{E}^{\nu_{k-1}^{\lambda}}[\lambda]$, $\tau_k = \mathbb{E}^{\nu_{k-1}^{\tau}}[\tau]$
4: Calculate ν_k^u by

$$\mathcal{C}_k^{-1} = \tau_k H^* H + \mathcal{C}_0(\lambda_k)^{-1}, \quad u_k = \mathcal{C}_k\big(\tau_k H^* d + \mathcal{C}_0(\lambda_k)^{-1} u_0\big).$$

5: Calculate ν_k^{λ} and ν_k^{τ} by

$$\nu_k^{\lambda} = Gamma(\tilde{\alpha}_0, \tilde{\beta}_0^k), \quad \nu_k^{\tau} = Gamma(\tilde{\alpha}_1, \tilde{\beta}_1^k),$$

where

$$\tilde{\alpha}_0 = \alpha_0 + \frac{K}{2}, \quad \tilde{\beta}_0^k = \beta_0 + \frac{1}{2}\mathbb{E}^{\nu_k^u}\bigg(\sum_{j=1}^{K}(u_j - u_{0j})^2 \alpha_j^{-1}\bigg),$$

$$\tilde{\alpha}_1 = \alpha_1 + \frac{N_d}{2}, \quad \tilde{\beta}_1^k = \beta_1 + \frac{1}{2}\mathbb{E}^{\nu_k^u}(\|Hu - d\|^2).$$

6: **until** $\max\big(\|u_k - u_{k-1}\|/\|u_k\|, \|\lambda_k - \lambda_{k-1}\|/\|\lambda_k\|, \|\tau_k - \tau_{k-1}\|/\|\tau_k\|\big) \le tol$
7: Return $\nu_k^u(du)\nu_k^{\lambda}(d\lambda)\nu_k^{\tau}(d\tau)$ as the solution.

where no explicit forms of H^*H and H^* need to be constructed. As demonstrated in [10, 43], the CG iterations may be terminated when sufficient reduction is made in the norm of the gradient and the prior operator may also be used to precondition the CG iterations. For the term $\mathbb{E}^{\nu_k^u}(\|Hu - d\|^2)$, by a straightforward generalization of the finite dimensional case [34] (Proposition 1.18 in [46] and (c) of Theorem VI.25 in [47] are used), we know that the core difficulty is to compute the following quantity

$$\mathrm{Tr}\big((\tau_k \mathcal{C}_0(\lambda_k)^{1/2} H^* H \mathcal{C}_0(\lambda_k)^{1/2} + Id)^{-1} \mathcal{C}_0(\lambda_k)^{1/2} H^* H \mathcal{C}_0(\lambda_k)^{1/2}\big), \tag{49}$$

where $\mathrm{Tr}(\cdot)$ denotes the operator trace. For a lot of practical applications, the operator H^*H is a compact operator. Then the analysis provided in Subsections 5.2 and 5.4 in [10] may applicable in the current setting, which implies that only a small number of eignvalues (independent of the dimension of the discretized parameter field) is required to be evaluated. We intend to investigate efficient implementations for large-scale problems in our future work.

3.2. Linear inverse problems with Laplace noise. As revealed by previous studies on low-rank matrix factorization [53], the Gaussian noise tends to be sensitive to outliers. Compared with the Gaussian distribution, the Laplace distribution is a *heavy-tailed* distribution that can better fit heavy noises and outliers. In this subsection, we develop VBM for the linear inverse problem (22) with the Laplace noise assumption.

For the noise vector $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_{N_d})^T \in \mathbb{R}^{N_d}$, we assume that each component ϵ_i follows the Laplace distribution with zero mean

$$\epsilon_i \sim \mathrm{Laplace}\bigg(0, \sqrt{\frac{\tau}{2}}\bigg) \tag{50}$$

with $\tau \in \mathbb{R}^+$. The probability density function of the above Laplace distribution is denoted by $p_L(\epsilon_i|0, \sqrt{\tau/2})$ that takes the following form:

$$p_L(\epsilon_i|0, \sqrt{\tau/2}) = \sqrt{\frac{2}{\tau}} \exp\Big(- \frac{|\epsilon_i|}{\sqrt{\tau/2}}\Big). \tag{51}$$

However, the Laplace distribution cannot be easily employed for posterior inference within the variational Bayes' inference framework [53]. A commonly utilized strategy will be employed to reformulate the Laplace distribution as a Gaussian scale mixture with exponential distributed prior to the variance, as discussed in [3, 53]. Let $p_E(z|\tau)$ be the density function of an exponential distribution, that is,

$$p_E(z|\tau) = \frac{1}{\tau} \exp\Big(- \frac{z}{\tau}\Big). \tag{52}$$

Then, we have

$$\begin{aligned} p_L\Big(x|0, \sqrt{\frac{\tau}{2}}\Big) =& \frac{1}{2}\sqrt{\frac{2}{\tau}} \exp\Big(- \sqrt{\frac{2}{\tau}}|x|\Big) \\ =& \int_0^\infty \frac{1}{\sqrt{2\pi z}} \exp\Big(- \frac{x^2}{2z}\Big) \frac{1}{\tau} \exp\Big(- \frac{z}{\tau}\Big) dz \\ =& \int_0^\infty p_N(x|0, z) p_E(z|\tau) dz. \end{aligned} \tag{53}$$

By substituting (50) into the above equation, we obtain

$$p_L(\epsilon_i|0, \sqrt{\tau/2}) = \int_0^\infty p_N(\epsilon_i|0, z_i) p_E(z_i|\tau) dz_i, \tag{54}$$

where $p_N(\epsilon_i|0, z_i)$ is the density function of a Gaussian measure on $\mathbb{R}$ with a zero mean and z_i variance. Thus, we can impose a two-level hierarchical prior instead of a single-level Laplace prior on each ϵ_i as

$$\epsilon_i \sim \mathcal{N}(0, z_i), \qquad z_i \sim \text{Exponential}(\tau). \tag{55}$$

Let $w_i = z_i^{-1}$. Given that $z_i \sim \text{Exponential}(\tau)$, we know that $w_i \sim \mu_0^{w_i}$ with $\mu_0^{w_i}$ being a probability distribution with the following probability density function:

$$\frac{1}{\tau} \exp\Big(- \frac{1}{\tau w_i}\Big) \frac{1}{w_i^2}. \tag{56}$$

Let W be a diagonal matrix with diagonal $w = \{w_1, w_2, \cdots, w_{N_d}\}$, and let

$$\mu_0^w = \prod_{i=1}^{N_d} \mu_0^{w_i}. \tag{57}$$

For the prior probability measure of u, similar to Subsection 3.1, we set this measure as for the Gaussian noise case, that is,

$$u \sim \mu_0^{u,\lambda} = \mathcal{N}(u_0, \mathcal{C}_0^K(\lambda)), \quad \lambda \sim \mu_0^\lambda = \text{Gamma}(\alpha_0, \beta_0). \tag{58}$$

By combining (57) and (58), we obtain the full prior probability measure as

$$\mu_0(du, d\lambda, dw) = \mu_0^{u,\lambda}(du)\mu_0^{\lambda}(d\lambda)\mu_0^{w}(dw). \tag{59}$$

For the reference probability measure, we set $\mu_r(du, d\lambda, dw) = \mu_r^u(du)\mu_r^\lambda(d\lambda)\mu_r^w(dw)$, where $\mu_r^u = \mathcal{N}(u_0, \mathcal{C}_0)$, $\mu_r^\lambda = \mu_0^\lambda$, and $\mu_r^w = \mu_0^w$. By similar calculations as shown in (34), we obtain

$$\Phi^0(u, \lambda, \tau) = \frac{1}{2}\sum_{j=1}^{K}(u_j - u_{0j})^2(\lambda - 1)\alpha_j^{-1} - \frac{K}{2}\log\lambda. \tag{60}$$

For the posterior probability measure, by assumptions on the noises (55)-(57), we have

$$\frac{d\mu}{d\mu_0}(u, \lambda, w) = \frac{1}{Z_\mu}|W|^{1/2}\exp\Big(-\frac{1}{2}\|W^{1/2}(Hu - d)\|^2\Big), \tag{61}$$

which implies $\Phi(u, \lambda, w) = \frac{1}{2}\|W^{1/2}(Hu - d)\|^2 - \frac{1}{2}\log|W|$. Similar to the Gaussian noise case, we specify the approximate probability measure as

$$\frac{d\nu}{d\mu_r}(u, \lambda, w) = \frac{1}{Z_r}\exp\Big(-\Phi_u^r(u) - \Phi_\lambda^r(\lambda) - \Phi_w^r(w)\Big). \tag{62}$$

With these preparations, we are ready to calculate the three potentials in (62). As discussed in Remark 15, we use $\lambda > 0$ as a hyper-parameter, which is not in accordance with our general theory. However, it can be made rigorous by considering $\lambda' = \ln\lambda$ as the hyper-parameter.

Calculate Φ_u^r: Following formula (20), we can derive

$$\begin{aligned}\Phi_u^r(u) =& \iint \frac{1}{2}\sum_{j=1}^{K}(u_j - u_{0j})^2(\lambda - 1)\alpha_j^{-1} + \frac{1}{2}\|W^{1/2}(Hu - d)\|^2 d\nu^\lambda d\nu^w + \text{Const} \\ =& \frac{\lambda^* - 1}{2}\sum_{j=1}^{K}\alpha_j^{-1}(u_j - u_{0j})^2 + \frac{1}{2}\|W^*(Hu - d)\|^2 + \text{Const},\end{aligned} \tag{63}$$

where $\lambda^* = \mathbb{E}^{\nu^\lambda}[\lambda]$ and $W^* = \text{diag}(\mathbb{E}^{\nu^w}[w_1], \mathbb{E}^{\nu^w}[w_2], \cdots, \mathbb{E}^{\nu^w}[w_{N_d}])$. From the equality (63), we easily conclude that

$$\frac{d\nu^u}{d\mu_r^u}(u) \propto \exp\Big(-\frac{1}{2}\|(W^*)^{1/2}(Hu - d)\|^2 - \frac{\lambda^* - 1}{2}\sum_{j=1}^{K}\alpha_j^{-1}(u_j - u_{0j})^2\Big), \tag{64}$$

which implies that u is distributed according to a Gaussian measure with a covariance operator and a mean value specified as $\mathcal{C}^{-1} = H^*W^*H + \mathcal{C}_0(\lambda^*)^{-1}$ and $u^* = \mathcal{C}\big(H^*W^*d + \mathcal{C}_0(\lambda^*)^{-1}u_0\big)$.

Calculate Φ_λ^r: Following formula (20), we can derive

$$\begin{aligned}\Phi_\lambda^r(\lambda) =& \iint \frac{1}{2}\sum_{j=1}^{K}(u_j - u_{0j})^2\alpha_j^{-1}\lambda - \frac{K}{2}\log\lambda d\nu^u d\nu^w + \text{Const} \\ =& \frac{1}{2}\mathbb{E}_{\nu^u}\Big(\sum_{j=1}^{K}(u_j - u_{0j})^2\alpha_j^{-1}\Big)\lambda - \frac{K}{2}\log\lambda + \text{Const}.\end{aligned} \tag{65}$$

Therefore, we have

$$\frac{d\nu^{\lambda}}{d\mu_r^{\lambda}}(\lambda) \propto \lambda^{K/2} \exp\Big(-\frac{1}{2}\mathbb{E}^{\nu^u}\Big(\sum_{j=1}^{K}(u_j - u_{0j})^2\alpha_j^{-1}\Big)\lambda\Big), \tag{66}$$

which implies that ν^{λ} is a Gamma distribution denoted by Gamma$(\tilde{\alpha}_0, \tilde{\beta}_0)$ with

$$\tilde{\alpha}_0 = \alpha_0 + K/2, \quad \tilde{\beta}_0 = \beta_0 + \frac{1}{2}\mathbb{E}^{\nu^u}\Big(\sum_{j=1}^{K}(u_j - u_{0j})^2\alpha_j^{-1}\Big). \tag{67}$$

Calculate Φ_w^r: Following formula (20), we derive

$$\begin{aligned}\Phi_w^r(w) =& \iint \frac{1}{2}\|W^{1/2}(Hu-d)\|^2 - \frac{1}{2}\log|W| d\nu^u d\nu^{\lambda} + \text{Const} \\ =& \frac{1}{2}\sum_{j=1}^{N_d}\mathbb{E}^{\nu^u}\big[(Hu-d)_i^2\big]w_i - \frac{1}{2}\sum_{j=1}^{N_d}\log w_i + \text{Const},\end{aligned} \tag{68}$$

which implies

$$\frac{d\nu^w}{d\mu_r^w}(w) \propto \prod_{j=1}^{N_d} w_j^{1/2}\exp\Big(-\frac{1}{2}\mathbb{E}^{\nu^u}\big[(Hu-d)_j^2\big]w_j\Big). \tag{69}$$

Because w is a finite dimensional random variable, we find

$$\begin{aligned} d\nu^w &\propto \prod_{j=1}^{N_d} w_j^{1/2}\exp\Big(-\frac{1}{2}\mathbb{E}^{\nu^u}\big[(Hu-d)_j^2\big]w_j\Big)\frac{1}{\tau}\exp\Big(-\frac{1}{\tau w_j}\Big)\frac{1}{w_j^2}dw \\ &\propto \prod_{j=1}^{N_d}\frac{1}{\tau w_j^{3/2}}\exp\Big(-\frac{1}{2}\mathbb{E}^{\nu^u}\big[(Hu-d)_j^2\big]w_j - \frac{1}{\tau w_j}\Big)dw.\end{aligned} \tag{70}$$

In other words, ν^w is an inverse Gaussian distribution denoted by $\prod_{j=1}^{N_d} IG(m_{w_j}, \zeta)$ with

$$m_{w_j} = \sqrt{\frac{2}{\tau\mathbb{E}^{\nu^u}\big[(Hu-d)_j^2\big]}}, \quad \zeta = \frac{2}{\tau}. \tag{71}$$

Specify the parameter τ: From (55), we know the parameter τ is directly related to noise variance parameter $z_i = w_i^{-1}$. Therefore, this parameter should be adjusted carefully to obtain reasonable results. Empirical Bayes [8] provides an off-the-shelf tool to be adaptively tuned based on the noise information extracted from the data by updating it through $\tau = \frac{1}{N_d}\sum_{j=1}^{N_d} m_{w_j}^{-1} + \zeta^{-1}$. Using this elaborate tool, τ can be properly adapted to real data variance.

Similar to the Gaussian noise case, an iterative algorithm, namely Algorithm 2, is constructed based on the above calculations. For large-scale problems, similar discussions of Algorithm 1 can be applied here. The only difference is that (49) is replaced by the following quantity:

$$\mathrm{Tr}\big((\tau_k\mathcal{C}_0(\lambda_k)^{1/2}H^*W_kH\mathcal{C}_0(\lambda_k)^{1/2} + Id)^{-1}\mathcal{C}_0(\lambda_k)^{1/2}H^*W_kH\mathcal{C}_0(\lambda_k)^{1/2}\big). \tag{72}$$

The quantity (72) can be calculated in the similar way as (49).

Algorithm 2 Variational approximation for the case of Laplace noise

1: Give an initial guess $\mu_0^{u,\lambda}$ (u_0 and λ), μ_0^{λ} (α_0 and β_0), μ_0^{w} and τ. Specify the tolerance *tol* and set $k = 1$.
2: **repeat**
3: Set $k = k + 1$
3: Calculate $\lambda_k = \mathbb{E}^{\nu_{k-1}^{\lambda}}[\lambda]$, $W_k = \mathrm{diag}\big(\mathbb{E}^{\nu^w}[w_1], \mathbb{E}^{\nu^w}[w_2], \cdots, \mathbb{E}^{\nu^w}[w_{N_d}]\big)$ and $\tau_k = \frac{1}{N_d}\sum_{j=1}^{N_d}(m_{w_j}^{k-1})^{-1} + (\zeta_{k-1})^{-1}$.
4: Calculate ν_k^u by

$$\mathcal{C}_k^{-1} = H^* W_k H + \mathcal{C}_0(\lambda_k)^{-1}, \quad u_k = \mathcal{C}_k\big(H^* W_k d + \mathcal{C}_0(\lambda_k)^{-1} u_0\big).$$

5: Calculate ν_k^{λ} and ν_k^{w} by

$$\nu_k^{\lambda} = Gamma(\tilde{\alpha}_0, \tilde{\beta}_0^k), \quad \nu_k^{w} = \prod_{j=1}^{N_d} IG(m_{w_j}^k, \zeta_k),$$

$$\tilde{\beta}_0^k = \beta_0 + \frac{1}{2}\mathbb{E}^{\nu_k^u}\left(\sum_{j=1}^{K}(u_j - u_{0j})^2 \alpha_j^{-1}\right), \quad \tilde{\alpha}_0 = \alpha_0 + K/2,$$

$$m_{w_j}^k = \sqrt{\frac{2}{\tau_k \mathbb{E}^{\nu_k^u}\left[(Hu - d)_j^2\right]}}, \quad \zeta_k = \frac{2}{\tau_k}.$$

6: **until** $\max\big(\|u_k - u_{k-1}\|/\|u_k\|, \|\lambda_k - \lambda_{k-1}\|/\|\lambda_k\|, \|\tau_k - \tau_{k-1}\|/\|\tau_k\|\big) \le tol$
7: Return $\nu_k^u(du)\nu_k^{\lambda}(d\lambda)\nu_k^w(dw)$ as the solution.

4. Concrete numerical examples.

4.1. Inverse source problem for Helmholtz equation. The inverse source problem (ISP) studied in this section is borrowed from [6, 7, 15, 30], which determines the unknown current density function from measurements of the radiated fields at multiple wavenumbers.

Consider the Helmholtz equation

$$\Delta v + \kappa^2(1 + q(x))v = u_s \quad \text{in} \quad \mathbb{R}^{N_s}, \tag{73}$$

where $N_s = 1, 2$ is the space dimension, κ is the wavenumber, v is the radiated scalar field, and the source current density function $u_s(x)$ is assumed to have a compact support. For the one-dimensional case, let the radiated field v satisfy the absorbing boundary condition: $\partial_r v = i\kappa v$. For the two-dimensional case, let the radiated field v satisfy the Sommerfeld radiation condition: $\partial_r v - i\kappa v = o(r^{-1/2})$ as $r = |x| \to \infty$. In addition, we employ an uniaxial perfect match layer (PML) technique to truncate the whole plane into a bounded rectangular domain when $N_s = 2$. For details on the uniaxial PML technique, see [5, 32] and references therein. Let D be the domain with absorbing layers, and Ω be the physical domain without absorbing layers.

The ISP aims to determine the source function u_s from the boundary measurements of the radiated field on the boundary $\partial\Omega$ for a series of wavenumbers. For clarity, we summarize the problem as follows:

Available data For $0 < \kappa_1 < \kappa_2 < \cdots < \kappa_{N_f} < \infty$ ($N_f \in \mathbb{N}^+$), and measurement points $x^1, x^2, \cdots, x^{N_m} \in \partial\Omega$, we denote

$$d^{\dagger} := \big\{v(x^i, \kappa_j)\,|\, i = 1, 2, \cdots, N_m, \text{ and } j = 1, 2, \cdots, N_f\big\}.$$

The available data set is $d := d^{\dagger} + \epsilon$, where ϵ is the measurement error.

Unknown function The source density function u_s needs to be determined.

Generally, we let $\mathcal{F}_\kappa$ be the forward operator that maps u_s to the solution v when the wavenumber is κ , and let $\mathcal{M}$ be the measurement operator mapping v to the available data. With these notations, the problem can be written abstractly as

$$d_\kappa = H_\kappa(u_s) + \epsilon_\kappa, \tag{74}$$

where $H_\kappa := \mathcal{M} \circ \mathcal{F}_\kappa$ is the forward operator, and ϵ_κ is the random noise.

To avoid inverse crime, we use a fine mesh to generate data and a rough mesh for the inversion. For the one-dimensional problem, meshes with mesh numbers of 1000 and 600 are used for the data generation and inversion, respectively. For the two-dimensional problem, we will provide details in the sequel.

When the dimension of the parameters is relatively low, the proposed Algorithms 1 and 2 are similar to the one build for the finite-dimensional case. Detailed comparisons with the MCMC algorithm have been given in [33, 34], which reflect that highly accurate inferences can be generated. Hence we will not present a comparison with the MCMC algorithm in the sequel for a relatively low dimensional case. For the infinite-dimensional Bayesian method with hyper-parameters, the noncentered algorithms are a more appropriate choice as illustrated in [1]. Using the proposed general framework for the noncentered parameterize strategy and providing a comparison with the method proposed in [1] could be an interesting future research problem.

It should be indicated that the finite element method is implemented by employing the open software FEniCS (Version 2018.1.0). For additional information on FEniCS, see [39]. All programs were run on a personal computer with Intel(R) Core(TM) i7-7700 at 3.60 GHz (CPU), 32 GB (memory), and Ubuntu 18.04.2 LTS (OS).

4.2. One-dimensional ISP. For clarity, we list the specific choices for some parameters introduced in Section 3 as follows:

- The operator $\mathcal{C}_0$ is chosen to be $(\mathrm{Id} - \partial_{xx})^{-1}$ and taken $\epsilon = 10^{-3}$. Here, the Laplace operator is defined on Ω with the zero Dirichlet boundary condition.
- The wavenumber series are specified as $\kappa_j = j$ with $j = \frac{1}{2}, 1, \frac{3}{2}, 2, \cdots, 50$.
- Let domain Ω be an interval $[0, 1]$, with $\partial\Omega = \{0, 1\}$. And the available data are assumed to be $\{v(x^i, \kappa_j) \,|\, i = 1, 2,\ x^1 = 0,\ x^2 = 1, \text{ and } j = 1, 2, \cdots, 100\}$.
- The initial values required by Algorithm 1 are chosen as $u_0 = 0, \alpha_0 = \alpha_1 = 1, \beta_0 = 10^{-1}, \beta_1 = 10^{-5}$. The initial values required by Algorithm 2 are chosen as $u_0 = 0, \alpha_0 = 1, \beta_0 = 10^{-1}, \tau = 10^{-7}$.
- The function $q(x)$ in the Helmholtz equation is taken to be constant zero.
- The ground truth source function u_s is defined as

$$u_s(x) = 0.5 \exp(-300(x - 0.4)^2) + 0.5 \exp(-300(x - 0.6)^2).$$

According to the studies presented in [38], for this simple one-dimensional case, we will not take a recursive strategy but combine instead all data together with the forward operator denoted by H and defined by $H = (H_{\kappa_1}, H_{\kappa_2}, \cdots, H_{\kappa_{100}})^T$. Based on these settings, we provide some basic theoretical properties of the prior and posterior sampling functions as follows:

- The prior probability measure for u_s is Gaussian with the covariance operator $\mathcal{C}_0^K(\lambda)$ with $\lambda \in \mathbb{R}^+$. According to Theorem 12 illustrated in [19], we know

that if u_s is drawn from the prior measure, and then the following holds

$$u_s \in W^{t,2}(\Omega) \quad \text{for } t < \frac{1}{2}, \quad \text{and} \quad u_s \in C^{0,t}(\Omega) \quad \text{for } t < \frac{1}{2},$$

where $W^{t,2}(\Omega)$ is the usual Sobolev space with t times derivative belonging to $L^2(\Omega)$, and $C^{0,t}$ is the conventional Hölder space.

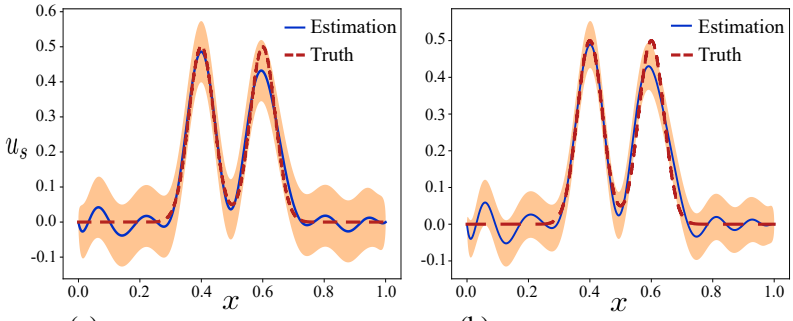
- For Algorithm 1, every posterior mean estimate u_k has the following form:

$$u_k = (\tau_k H^* H + \mathcal{C}_0(\lambda_k)^{-1})^{-1} \tau_k H^* d.$$

Given that H maps a function in $L^2(\Omega)$ to $\mathbb{R}^{200}$, we know that $H^* d$ is at least a function belonging to L^2. Considering the specific choices of $\mathcal{C}_0$, we have $u_k \in W^{2,2}(\Omega)$. For Algorithm 2, we can derive similar conclusions.

REMARK 16. *By employing the "Bayesianize-then-discretize" method, we can analyze the prior and posterior sampling functions rigorously. It is one of the advantages of employing our proposed infinite-dimensional VBM.*

Gaussian noise case: Let $d^\dagger$ be the data without noise. Then, we construct noisy data by setting $d = d^\dagger + \sigma\xi$ with $\sigma = 10^{-3}$ and ξ is a random variable sampled from the standard normal distribution.



(a) Estimation obtained by Algorithm 1 (b) Estimation obtained by Algorithm 2

FIG. 1. *The truth and estimated functions when the data are polluted by Gaussian noise. (a): the estimated function obtained by Algorithm 1 is denoted by the blue solid line, and the truth is denoted by the red dashed line; (b): the estimated function obtained by Algorithm 2 is denoted by the blue solid line, and the truth is denoted by the red dashed line. In both plots the shaded areas represent the pointwise mean plus and minus two standard deviations from the mean (corresponding roughly to the 95% confidence region).*

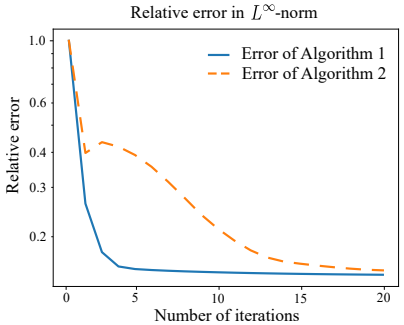


FIG. 2. *Relative errors of the estimated means in the L^∞-norm of Algorithms 1 and 2.*

In Figure 1, we depict the truth and estimated sources obtained by Algorithms 1 and 2, respectively. Visually, both algorithms provide reasonable results. In addition, we demarcate the 95% confidence region by the shaded area to display the uncertainties estimated by these two algorithms. The truth falls entirely into the confidence region given by Algorithm 1, and the truth lies mostly within the confidence region given by Algorithm 2. This may indicate that for the Gaussian noise case, Algorithm 1 can provide a more reliable estimation, which is in accordance with our assumptions.

To give a more elaborate comparison, we present the relative errors of the estimated means in the L^∞-norm of the two algorithms in Figure 2. The relative error of the conditional mean estimate used here is defined as follows

$$\text{relative error} = \|u - u_s\|_{L^\infty} / \|u_s\|_{L^\infty},$$

where u is the estimated function generated by our algorithm and u_s is the true source function. The blue solid line and orange dashed line denote the relative errors obtained by Algorithms 1 and 2, respectively. Obviously, these two algorithms can provide comparable results after convergence. However, Algorithm 1 converges much faster than Algorithm 2, which is reasonable because the weight parameters used for detecting impulsive noises may reduce the convergence speed.

The parameter τ given by Algorithm 1 provides an estimate of the noise variance through $\sigma = \sqrt{\tau^{-1}}$. The true value of σ is 0.001 in our numerical example. To generate a repeatable results, we specify the random seeds in numpy to some certain numbers by numpy.random.seed(i) with i specified as some designated integers. The estimated σ is equal to $0.000953, 0.001101, 0.001022, 0.001003$, and 0.001041 when the random seeds are specified as $1, 2, 3, 4$, and 5, respectively, thereby illustrating the effectiveness of our proposed algorithm.

Laplace noise case: As for the Gaussian noise case, let $d^\dagger$ be the noise-free measurement. The noisy data are generated as follows:

$$d_i = \begin{cases} d_i^\dagger, & \text{with probability } 1 - r, \\ d_i^\dagger + \epsilon\xi, & \text{with probability } r, \end{cases}$$

where ξ follows the uniform distribution $U[-1, 1]$, and (ϵ, r) controls the noise pattern, r is the corruption percentage, and ϵ is the corruption magnitude. In the following, we take $r = 0.5$ and $\epsilon = 0.1$. We plot the clean and noisy data in Figure 3, which illustrates that the clean data are heavily polluted.

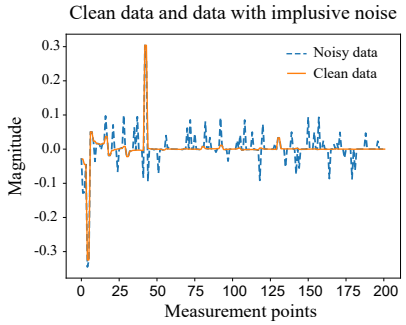
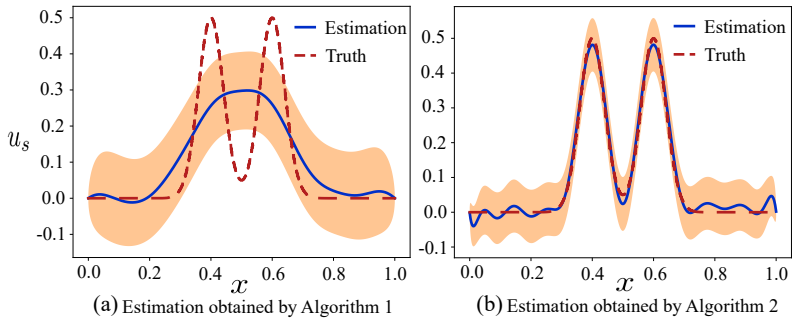
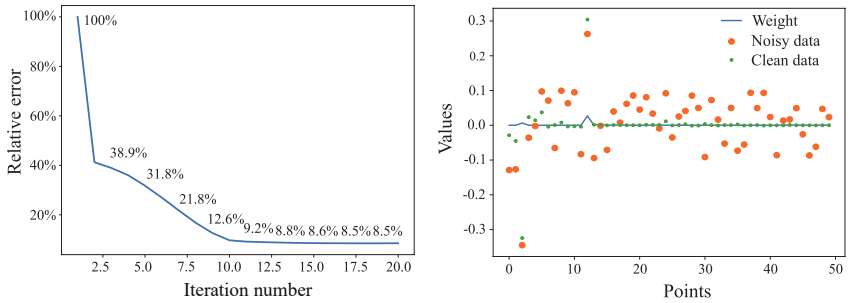


FIG. 3. *Clean and noisy data. The orange solid line represents the clean data, and the blue dashed line represents the data with impulsive noise.*



(a) Estimation obtained by Algorithm 1 (b) Estimation obtained by Algorithm 2

FIG. 4. *The truth and estimated functions when the data are polluted by impulsive noise. (a): The estimated function obtained by Algorithm 1 is denoted by the blue solid line, and the truth is denoted by the red dashed line; (b): The estimated function obtained by Algorithm 2 is denoted by the blue solid line, and the truth is denoted by the red dashed line. The shaded areas in both panels represent the pointwise mean plus and minus two standard deviations from the mean (corresponding roughly to the 95% confidence region).*



(a) Relative error in L^∞-norm (b) Part of the weights, noisy and clean data

FIG. 5. *(a): Relative errors in the L^∞-norm obtained by Algorithm 2; (b): Weight, noisy, and clean data at the data points with impulsive noise (only points with impulsive noise, not all points).*

In Figure 4, we show the estimated functions obtained by Algorithms 1 and 2 in the left and right panels, respectively. Obviously, based on the Gaussian noise assumption, Algorithm 1 cannot provide a reasonable estimate, and the estimated confidence region may be unreliable. However, based on the Laplace noise assumption, Algorithm 2 provides an accurate estimate. Given that Algorithm 1 fails to converge to a reasonable estimation, we only provide the relative errors in the L^∞-norm of Algorithm 2 on the left panel of Figure 5. From these relative errors, we can find that Algorithm 2 rapidly converges even if the data are heavily polluted by noise. The right panel of Figure 5 plots the noisy and clean data points at those data points where noises are added. We plot the weight vector at the corresponding data points. From this figure, we can clearly see that the elements of the weight vector are all with small values, which is in accordance with our theory. The weight vectors at the noisy data points are adjusted to small values during the iteration. This reveals the outlier removal mechanism of Algorithm 2.

4.3. Two-dimensional ISP. In this subsection, we solve the two-dimensional ISP. Directly computing the covariance operator for the two-dimensional problem is difficult due to the large memory requirements and computational inefficiency. Here, we employ a simple method that employs a rough mesh approximation to compute

the covariance. The source function u_s can be expanded under basis functions as follows:

$$u_s(x) = \sum_{i=1}^{\infty} u_{si}\varphi_i(x). \tag{75}$$

Given that these basis functions can be taken as the finite element basis, the source function can be approximated as

$$u_s(x) \approx \sum_{i=1}^{N_t} u_{si}\varphi_i(x). \tag{76}$$

The covariances involved in Algorithms 1 and 2 are all computed by taking a small N_t in (76). For many applications such as medical imaging, we may compute the operator H^*H (not depending on the source function) with a small N_t before the inversion. To evaluate accurately as the wavenumber increases, we compute the mean function by gradient descent with a fine mesh discrete PDE solver and then project the source function to the rough mesh for computing variables relying on the covariance operators.

Unlike the one-dimensional case, we employ the sequential method used in [6] that provides a more stable recovery for multi-frequency inverse problems. Specifically, for $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_{N_f} < \infty$ and each problem $d_{\kappa_i} = H_{\kappa_i}(u_s) + \epsilon_{\kappa_i} (i = 1, \cdots, N_f)$, we assume the prior measure is $\mu_{0i}^{u,\lambda} = \mathcal{N}(\bar{u}_{i-1}, \mathcal{C}_0^K(\lambda))$ with $\bar{u}_{i-1}$ denoting the conditional mean estimate when the wavenumber is κ_{i-1} ($\bar{u}_0$ is assumed to be some initial guess u_s^0). For the Gaussian noise case with $i = 1, 2, \cdots, N_f$, we have the following Bayesian formula

$$\frac{d\mu^i}{d\mu_{0i}}(u, \lambda, \tau) \propto \exp\Big(-\frac{\tau}{2}\|H_{\kappa_i}(u) - d_{\kappa_i}\|^2\Big), \tag{77}$$

where $\mu_{0i}(du, d\lambda, d\tau) = \mu_{0i}^{u,\lambda}(du)\mu_0^{\lambda}(d\lambda)\mu_0^{\tau}(d\tau)$ with $\mu_0^{\lambda}, \mu_0^{\tau}$ are defined as in Subsection 3.1 and μ^i is the posterior measure when wavenumber is equal to κ_i. The posterior measure $\mu^{\kappa_{N_f}}$ will be employed to quantify the uncertainties of the final estimate. For a similar sequential formulation as above, we refer to Subsection 6.4.1 in [38]. It is not hard to formulate a sequential approach for the Laplace noise case. The details are omitted for conciseness. The iteration details are presented in Algorithm 3, in which $\|\cdot\|_{\mathcal{C}_0^K(\lambda)}$ denotes the Cameron-Martin norm corresponding to the Gaussian measure $\mathcal{N}(0, \mathcal{C}_0^K(\lambda))$. In the following, when we say that Algorithm 1 is employed, we actually means that Algorithm 3 is employed in combination with Algorithm 1. Similarly, when we say that Algorithm 2 is employed, we mean that Algorithm 3 is employed in combination with Algorithm 2.

REMARK 17. *It should be pointed out that the simple "rough mesh approximation" method employed in Algorithm 3 is only applicable to problems with a simple form (e.g., a localized source) on simple geometry. This method is not suitable for dealing with more complex problems in three-dimension or even in two-dimension where a large N_t is needed (e.g., high-resolution recovery with data of high wavenumbers). Our aim is to give an illustration of the proposed method. For more advanced techniques designed for large-scale problems, [10] can be referred to, which provides a scalable approach for the infinite-dimensional Bayesian approach with linear approximations.*

Algorithm 3 VBM for two-dimensional ISP with multi-frequencies

1: Give an initial guess of the unknown source u_s, denoted by u_s^0.
2: For i from 1 to N_f (iterate from low wavenumber to high wavenumber)
3: Specify the prior measure of u_s as $\mu_{0i}^{u,\lambda} = \mathcal{N}(u_s^{i-1}, \mathcal{C}_0^K(\lambda))$. Running iterations of Algorithms 1 or 2 for k until some stopping criterion is satisfied. For $k = 1$, rough approximate of H and source is employed; For $k > 1$, the gradient descent method is employed to solve

$$u_s^k := \arg\min_{u_s} \frac{\tau_k}{2} \|H_{\kappa_i}(u_s) - d_{\kappa_i}\|^2 + \|u_s - u_s^{i-1}\|^2_{\mathcal{C}_0^K(\lambda_k)},$$

which generate a conditional mean estimate on a fine mesh. In all of the iterations, rough approximate Hessian has been used to update distributions of hyper-parameters λ, τ (Algorithm 1) or w (Algorithm 2).
5: End for
6: Return the approximate probability measure ν.

The fully nonlinear case has been investigated by using a stochastic Newton MCMC method in [43]. Then, Metropolize-then-discretize and discretize-then-Metropolize have been analyzed carefully for large-scale problems [12]. In 2019, an approximate sampling method based on some randomized MAP estimates has been investigated in detail [50]. All these studies provide valuable ideas of designing algorithms of large-scale inverse problems. For more studies in this direction, we refer to [11, 28, 40].

REMARK 18. *In Algorithm 3, we use approximations on a rough mesh for the first iteration of every wavenumber, which may provide an initial inaccurate adjustment for the parameters employed in Algorithms 1 and 2. In our numerical experiments, we only take three iterations for the third step to obtain an estimation.*

REMARK 19. *To employ sampling-type methods such as the MCMC algorithm, researchers often parameterize the unknown source function carefully to reduce the dimension, e.g., assume that the sources are point sources, then parameterize the source function by numbers, locations, and amplitudes [22]. For employing MCMC algorithm [16, 23] in our setting, the computational complexity is unacceptable for two reasons: Calculation with many wavenumbers are needed for multi-frequency problems and a large number of samples need to be generated for each wavenumber; For each problem (77), we did not assume any parametric form of the source function which makes the parameters of source equal to the dimension of the discretization (much more parameters than the usually used parametric form). However, the proposed Algorithm 3 only takes several times of computational time compared with the classical iterative algorithms [6, 7, 29] to provide estimations of uncertainties.*

Before going further, we list the specific choices for some parameters introduced in Section 3 as follows:

- The operator $\mathcal{C}_0$ is chosen as $(-\Delta + \mathrm{Id})^{-2}$. Here, the Laplace operator is defined on Ω with the zero Dirichlet boundary condition.
- Take the discrete truncate level $N_t = 1681$ and the number of measurement points $N_m = 200$. The basis functions $\{\varphi_j\}_{j=1}^{\infty}$ are specified as second-order finite element basis functions.
- For Algorithm 3 combined with Algorithm 1, the wavenumber series are specified as $\kappa_j = j$ with $j = 1, 3, 5, \cdots, 35$. For Algorithm 3 combined with Algo-

rithm 2, the wavenumber series are specified as $\kappa_j = j$ with $j = 1, 2, 3, \cdots, 35$.

- The scatterer function $q(x)$ is defined as follows:

$$\begin{aligned} q(x_1, x_2) =& 0.3(4 - 3x_1)^2 e^{\left(-9(x_1-1)^2 - 9(x_2-2/3)^2\right)} \\ & - \left(0.6(x_1 - 1) - 9(x_1 - 1)^3 - 3^5(x_2 - 1)^5\right) e^{\left(-9(x_1-1)^2 - 9(x_2-1)^2\right)} \\ & - 0.03e^{-9(x_1-2/3)^2 - 9(x_2-1)^2}, \end{aligned}$$

which is the function used in Subsection 2.6 in [6].

- The true source function u_s is defined as follows:

$$u_s(x) = 0.5e^{-100((x_1-0.7)^2+(x_2-1)^2)} + 0.3e^{-100((x_1-1.3)^2+(x_2-1)^2)}.$$

- To avoid the inverse crime, a mesh with mesh number 125000 is employed for generating the data. For the inversion, two types of meshes are employed: a mesh with mesh number 28800 is employed when the wavenumbers are below 20, and a mesh with mesh number 41472 is employed when the wavenumbers are greater than 20.

The case of Gaussian noise: Let $d^\dagger$ be the data without noise. The synthetic noisy data d are generated by $d_j = d^\dagger + \sigma\xi_j$, where $\sigma = \max_{1\le j\le N_m}\{|d_j^\dagger|\}L_{\text{noise}}$ with L_{noise} denoting the relative noise level and ξ_j denoting the standard normal random variables. In our experiments, we take $L_{\text{noise}} = 0.05$, that is 5% of noises are added.

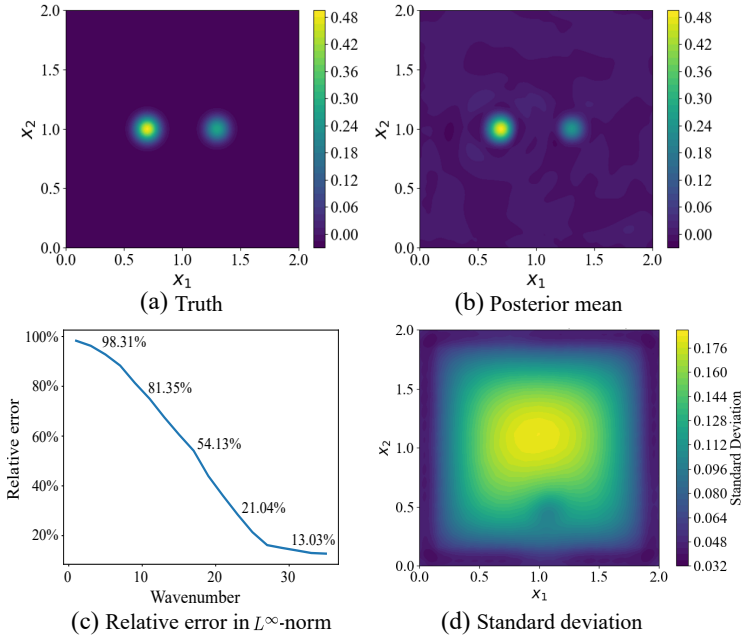


FIG. 6. *(a): The true source function; (b): The posterior mean estimate obtained by Algorithm 1; (c): Relative error of the estimated means in L^∞-norm obtained by Algorithm 1; (d): Estimated standard deviation obtained by Algorithm 1.*

In Figure 6, we show the inference results obtained by Algorithm 1. We show the true source function on the top left and the posterior mean estimate on the top right. Visually, the estimate is similar to the truth, and only some small fluctuations

in the background are observed. In the bottom left, we show the relative errors of the estimated means obtained by Algorithm 1 as the wavenumber increases, which is in accordance with the results obtained by classical iterative approaches. In the bottom right, we show the estimated standard deviation obtained by Algorithm 1 that quantifies the uncertainties of the posterior mean estimation. We see that the uncertainties are small on the boundary where data are collected. The areas with the largest uncertainties are in the middle, which is a reasonable result since that area can be recovered only when data generated by high wavenumbers are employed.

The case of Laplace noise: For the Laplace noise case, let $d^\dagger$ be the noise-free measurement. The noisy data are generated as

$$d_i = \begin{cases} d_i^\dagger, & \text{with probability } 1-r, \\ d_i^\dagger + \epsilon\xi, & \text{with probability } r, \end{cases}$$

where ξ follows the uniform distribution $U[-1,1]$, (ϵ, r) controls the noise pattern, r is the corruption percentage, and ϵ is the corruption magnitude defined by $\epsilon = \max_{1\le j\le N_m}\{|d_j^\dagger|\}L_{\text{noise}}$ with L_{noise} denoting the relative noise level. In our experiments, we take $L_{\text{noise}} = 1$ and $r = 0.2$ or 0.5.

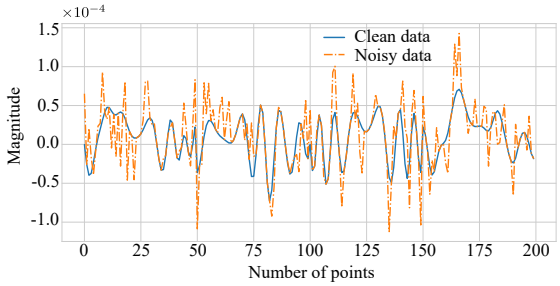


FIG. 7. *Clean and noisy data obtained when the wavenumber is* 34. *The blue solid line represents the clean data, and the dashed orange line represents the noisy data with* $r = 0.5$.

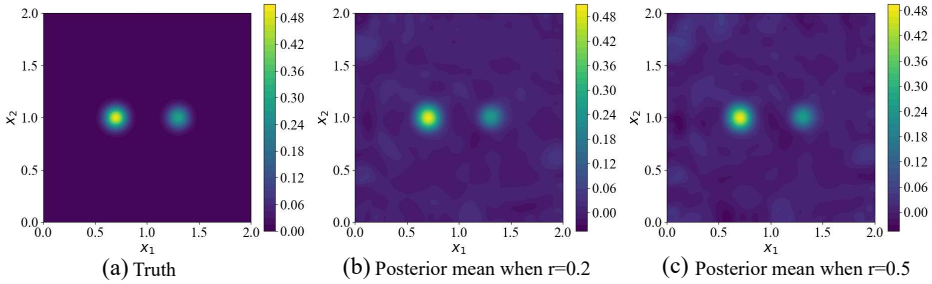
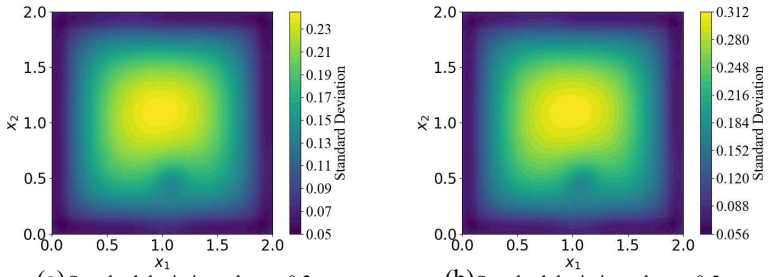


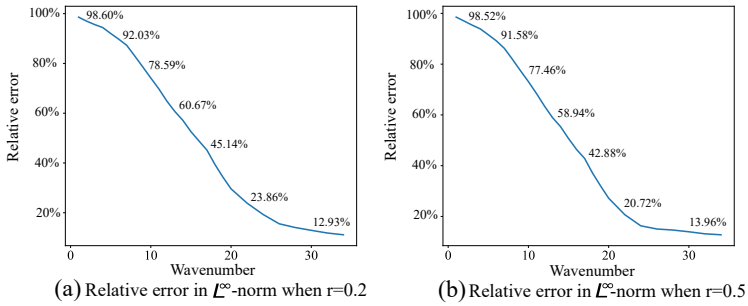
FIG. 8. *(a): The true source function; (b): The posterior mean estimate provided by Algorithm 2 from noisy data with* $r = 0.2$ *(20% of data are polluted); (c): The posterior mean estimate provided by Algorithm 2 from noisy data with* $r = 0.5$ *(50% of data are polluted).*

The noisy and clean data when the wavenumber is 34 and $r = 0.5$ are shown in Figure 7. Obviously, the data are heavily contaminated by noise. Figure 8 shows the true source function and the posterior mean estimates generated by Algorithm 2 when $r = 0.2$ and $r = 0.5$ on the left, middle, and right panels, respectively. No essential



(a) Standard deviation when r=0.2 (b) Standard deviation when r=0.5

FIG. 9. *Standard deviation of the numerical solution obtained by Algorithm 3 combined with Algorithm 2. (a): Estimated standard deviation when $r = 0.2$ (20% of data are polluted); (b): Estimated standard deviation when $r = 0.5$ (50% of data are polluted).*



(a) Relative error in L^∞-norm when r=0.2 (b) Relative error in L^∞-norm when r=0.5

FIG. 10. *Relative errors of the estimated means in L^∞-norm of Algorithm 3 combined with Algorithm 2. (a): Relative errors for $r = 0.2$; (b): Relative errors for $r = 0.5$.*

differences can be observed between the posterior mean estimates when $r = 0.2$ and $r = 0.5$. However, the Bayes' method not only provides point estimates (e.g., posterior mean estimates) but also delivers the reliability of the obtained estimations. Figure 9 shows the standard deviations provided by Algorithm 2 when $r = 0.2$ and $r = 0.5$ on the left and right panels, respectively. The standard deviations are smaller when $r = 0.2$, which is reasonable given that 80% of the data are clean and only 50% of the data are clean when $r = 0.5$. Figure 10 shows the relative errors in L^∞-norm obtained by Algorithm 2 with $r = 0.2, 0.5$ on the left and right panels, respectively. Under both settings, the relative errors of the posterior mean estimates rapidly decrease.

REMARK 20. *The wavenumber series in the present paper are not chosen carefully in an optimal way. There are some studies focused on the strategies for selecting appropriate wavenumbers to give an accurate estimate under the framework of regularization methods for geophysical inverse problems [42]. Here, we choose more wavenumbers for the Laplace noise model based on a simple intuitive idea. More data are required when more hyper-parameters need to be inferred (The Laplace noise model has more parameters than the Gaussian noise model).*

5. Conclusion. In this paper, we have generalized the finite-dimensional mean-field approximate based variational Bayes' method (VBM) to infinite-dimensional space, which provides a mathematical foundation for applying VBM to the inverse problems of PDEs. A general theory for the existence of minimizers has been established, and by introducing the concept of reference probability measure, the mean-field approximate theory has been constructed for functions. The established general theo-

877 ry is then applied to abstract linear inverse problems with Gaussian and Laplace noise
878 assumptions. Numerical examples for the inverse source problems of Helmholtz equa-
879 tions are investigated in details to highlight the effectiveness of the proposed theory
880 and algorithms.
881 There are numerous interesting problems that are worthy of being further in-
882 vestigated. Introducing a more reasonable setting of the intrinsic dimension will be
883 important. The recently published paper [14] provides some promising ideas. For the
884 infinite-dimensional Bayesian method with hyper-parameters, noncentered parame-
885 terization [1] could be a more appropriate choice. Using the proposed theory under
886 the noncentered parameterization is a problem worthy of further investigation.

887 **Acknowledgments.** The authors would like to thank the anonymous referees
888 for their comments and suggestions, which helped to improve the paper significantly.
889 We also thank Ms. Ying Feng for her thorough polishing of this paper. This work was
890 partially supported by the NSFC under grant Nos. 11871392, 62076196, 11690011,
891 61721002, U1811461, and the key project of NSFC under grant No. 12031003.

892 REFERENCES

893 [1] S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos, and A. M. Stuart, *Analysis of the Gibbs*
894 *sampler for hierarchical inverse problems*, SIAM/ASA J. Uncertainty Quantification, 2
895 (2014), pp. 511–544.
896 [2] S. Agapiou, M. Burger, M. Dashti, and T. Helin, *Sparsity-promoting and edge-preserving*
897 *maximum a posteriori estimators in non-parametric Bayesian inverse problems*, Inverse
898 Probl., 34 (2018), p. 045002.
899 [3] D. F. Andrews and C. L. Mallows, *Scale mixtures of normal distributions*, J. R. Stat. Soc.
900 B, 36 (1974), pp. 99–102.
901 [4] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, *Solving inverse problems using*
902 *data-driven models*, Acta Numer., 28 (2019), pp. 1–174.
903 [5] G. Bao, S. N. Chow, P. Li, and H. Zhou, *Numerical solution of an inverse medium scattering*
904 *problem with a stochastic source*, Inverse Probl., 26 (2010), p. 074014.
905 [6] G. Bao, P. Li, J. Lin, and F. Triki, *Inverse scattering problems with multi-frequencies*, Inverse
906 Probl., 31 (2015), p. 093001.
907 [7] G. Bao, S. Lu, W. Rundell, and B. Xu, *A recursive algorithm for multi-frequency acoustic*
908 *inverse source problems*, SIAM J. Numer. Anal., 53 (2015), pp. 1608–1628.
909 [8] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, New York, NY,
910 USA, 2006.
911 [9] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart, *Convergence rates of general regu-*
912 *larization method for statistical inverse problems and applications*, SIAM J. Numer. Anal.,
913 45 (2007), pp. 2610–2636.
914 [10] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for*
915 *infinite-dimensional Bayesian inverse problems part I: The linearized case, with applica-*
916 *tion to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523.
917 [11] T. Bui-Thanh and M. A. Girolami, *Solving large-scale PDE-constrained Bayesian inverse*
918 *problems with Riemann manifold Hamiltonian Monte Carlo*, Inverse Probl., 30.
919 [12] T. Bui-Thanh and Q. P. Nguyen, *FEM-based discretization-invariant MCMC methods for*
920 *PDE-constrained Bayesian inverse problems*, Inverse Probl. Imag., 10 (2016), pp. 943–975.
921 [13] M. Burger and F. Lucka, *Maximum a posteriori estimates in linear inverse problems with*
922 *log-concave priors are proper Bayes estimators*, Inverse Probl., 30 (2014), p. 114004.
923 [14] P. Chen, K. Wu, J. Chen, T. OLeary-Roseberry, and O. Ghattas, *Projected Stein vari-*
924 *ational Newton: A fast and scalable Bayesian inference method in high dimensions*, in
925 Advances in Neural Information Processing Systems 32, 2019, pp. 15130–15139.
926 [15] J. Cheng, V. Isakov, and S. Lu, *Increasing stability in the inverse source problem with many*
927 *frequencies*, J. Differ. Equations, 260 (2016), pp. 4786–4804.
928 [16] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions:*
929 *modifying old algorithms to make them faster*, Stat. Sci., 28 (2013), pp. 424–446.
930 [17] T. Cui, K. J. H. Law, and Y. M. Marzouk, *Dimension-independent likelihood-informed*
931 *MCMC*, J. Comput. Phys., 304 (2016), pp. 109–137.

[18] M. Dashti, K. J. Law, A. M. Stuart, and J. Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Probl., 29 (2013), p. 095017.
[19] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, Handbook of Uncertainty Quantification, (2017), pp. 311–428.
[20] M. M. Dunlop, M. A. Iglesias, and A. M. Stuart, *Hierarchical Bayesian level set inversion*, Stat. Comput., 27 (2017), pp. 1555–1584.
[21] M. M. Dunlop and A. W. Stuart, *MAP estimators for piecewise continuous inversion*, Inverse Probl., 32 (2016), p. 105003.
[22] S. Engel, D. Hafemeyer, C. Münch, and D. Schaden, *An application of sparse measure valued Bayesian inverse to acoustic sound source identification*, Inverse Probl., 35 (2019), p. 075005.
[23] Z. Feng and J. Li, *An adaptive independence sampler MCMC algorithm for Bayesian inferences of functions*, SIAM J. Sci. Comput., 40 (2018), pp. A1310–A1321.
[24] A. Fichtner, *Full Seismic Waveform Modelling and Inversion*, Springer, 2011.
[25] G. H. Golub and C. F. Van Load, *Matrix Computations*, 4 ed., 2014.
[26] N. Guha, X. Wu, Y. Efendiev, B. Jin, and B. K. Malick, *A variational Bayesian approach for inverse problems with skew-t error distribution*, J. Comput. Phys., 301 (2015), pp. 377–393.
[27] T. Helin and M. Burger, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Probl., 31 (2015), p. 085009.
[28] T. Isaac, N. Petra, G. Stadler, and O. Ghattas, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, J. Comput. Phys., 296 (2015), pp. 348–368.
[29] V. Isakov and S. Lu, *Increasing stability in the inverse source problem with attenuation and many frequencies*, SIAM J. Appl. Math., 78 (2018), pp. 1–18.
[30] V. Isakov and S. Lu, *Inverse source problems without (pseudo) convexity assumptions*, Inverse Probl. Imag., 12 (2018), pp. 955–970.
[31] K. Ito and B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*, World Scientific, 2015.
[32] J. Jia, B. Wu, J. Peng, and J. Gao, *Recursive linearization method for inverse medium scattering problems with complex mixture Gaussian error learning*, Inverse Probl., 35 (2019), p. 075003.
[33] B. Jin, *A variational Bayesian method to inverse problems with implusive noise*, J. Comput. Phys., 231 (2012), pp. 423–435.
[34] B. Jin and J. Zou, *Hierarchical Bayesian inference for ill-posed problems via variational method*, J. Comput. Phys., 229 (2010), pp. 7317–7343.
[35] J. P. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer Science & Business Media, Berlin, 2005.
[36] M. Lassas and S. Siltanen, *Can one use total variation prior for edge-preserving Bayesian inversion?*, Inverse Probl., 20 (2004), p. 1537.
[37] P. D. Lax, *Functional Analysis*, Wiley-Interscience, 2002.
[38] S. W. X. Lim, *Bayesian inverse problems and seismic inversion*, PhD thesis, University of Oxford, 2016.
[39] A. Logg, K. A. Mardal, and G. N. Wells, *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012.
[40] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, *A stochastic newton mcmc method for large-scale statistical inverse problems with application to seismic inversion*, SIAM J. Sci. Comput., 34 (2012), pp. A1460–A1487.
[41] A. G. D. G. Matthews, *Scalable Gaussian process inference using variational methods*, PhD thesis, University of Cambridge, 9 2016.
[42] G. Pan, L. Liang, and T. M. Habashy, *A numerical study of 3D frequency-domain elastic full-waveform inversion*, Geophysics, 84 (2019), pp. R99–R108.
[43] N. Petra, J. Martin, G. Stadler, and O. Ghattas, *A computational framework for infinite-dimensional Bayesian inverse problems, part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1525–A1555.
[44] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber, *Algorithms for Kullback-Leibler approximation of probability measures in infinite dimensions*, SIAM J. Sci. Comput., 37 (2015), pp. A2733–A2757.
[45] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber, *Kullback-Leibler approximation for probability measures on infinite dimensional space*, SIAM J. Math. Anal., 47 (2015), pp. 4091–4122.

[46] G. D. Prato, *An Introduction to Infinite-Dimensional Analysis*, Springer, 2006.
[47] M. Reed and B. Simon, *Functional Analysis I: Methods of Modern Mathematical Physics*, Elsevier (Singapore) Pte Ltd, revised and enlarged edition ed., 2003.
[48] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[49] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, 2005.
[50] K. Wang, T. Bui-Thanh, and O. Ghattas, *A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems*, SIAM J. Sci. Comput., 40 (2018), pp. A142–A171.
[51] K. Yang, N. Guha, Y. Efendiev, and B. K. Mallick, *Bayesian and variational Bayesian approaches for flows in heterogeneous random media*, J. Comput. Phys., 345 (2017), pp. 275–293.
[52] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, *Advances in variational inference*, IEEE T. Pattern Anal., (2018), pp. 1–1.
[53] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and Y. Yan, *l_1-norm low-rank matrix factorization by variational Bayesian method*, IEEE T. Neur. Net. Lear., 26 (2015), pp. 825–839.