A Variational Bayesian Approach for Inverse Problems with Skew-t Error Distributions

Nilabja Guha[1,4], Xiaoqing Wu[2], Yalchin Efendiev[1], Bangti Jin[3], Bani K. Mallick[2]

Abstract

In this work, we develop a novel robust Bayesian approach to inverse problems with data errors following a skew-t distribution. A hierarchical Bayesian model is developed in the inverse problem setup. The Bayesian approach contains a natural mechanism for regularization in the form of a prior distribution, and a LASSO type prior distribution is used to strongly induce sparseness. We propose a variational type algorithm by minimizing the Kullback-Leibler divergence between the true posterior distribution and a separable approximation. The proposed method is illustrated on several two-dimensional linear and nonlinear inverse problems, e.g. Cauchy problem and permeability estimation problem.

Keywords: Bayesian inverse problems; hierarchical Bayesian model; variational approximation; Kullback-Leibler divergence

1. Introduction

Mathematical models are frequently used in science and engineering, with applications in weather forecasting, climate prediction, chemical kinetics and oil reservoir forecasting. In these mathematical models, there are often model parameters or inputs that have to be estimated from indirect observational data, which constitutes an inverse problem. In practice, observations are inevitably noisy, due to limited precision of measurement sensors. Often the noises exhibit both heavy tail and skewness behavior, hence flexible non-Gaussian distributions are needed to adequately accommodate these features and to fully extract all relevant information. Further, inverse problems are often ill-posed in the sense that the solution lacks a stable dependence on data perturbations, which necessitates the use of regularization techniques [16]. Hence, obtaining a stable and accurate numerical solution is generally a daunting task.

In this work, we shall develop a robust hierarchical Bayesian model which provides a principled yet very flexible framework for solving inverse problems. We incorporate regularization through a suitable prior distribution. Moreover, we allow a heavy-tailed distribution for the error via the likelihood function. The posterior distribution is obtained by using Bayes' theorem. In this way, it yields an ensemble of inverse solutions consistent with the

[1] Department of Mathematics, Texas A&M University, College Station, TX 77843, USA
[2] Department of Statistics, Texas A&M University, College Station, TX 77843, USA
[3] Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
[4] Email of the corresponding author: nguha@math.tamu.edu

given data to various extents. In particular, it enables uncertainty quantification of a specific inverse solution within the ensemble. Furthermore, it provides a flexible regularization technique by selecting nuance parameters, e.g., regularization parameter and noise level, adaptively and automatically, through hierarchical Bayesian modeling, via e.g., the full or empirical Bayesian treatment.

Inference based on Bayesian hierarchical models provides an attractive tool for solving inverse problems due to its inherent ability to jointly estimate the regularizing parameters, noise level and inverse solution as well as to calibrate their uncertainties. The Gaussian error model is the most popular tool used in the existing Bayesian inverse problem setup. However, in practice, the normality assumption is usually violated because of the presence of skewness and kurtosis in real data [12]. Thus, one may seek more flexible parametric families that are capable of capturing such features of the data. The family of skew-normal distributions to capture the skewness in the data has been widely studied due to its mathematical tractability and appealing probabilistic properties [2, 6, 4, 3]. One further extension of the skew-normal distribution is the skew-t distribution which allows both nonzero skewness and heavy tails in the distribution [8]. For a general background on the skew-normal and related distributions, see [15] for an overview.

Markov chain Monte Carlo (MCMC) methods work particularly well in this setup and is the major engine that has fueled the development and application of Bayesian hierarchical models [14]. Despite the popularity of MCMC based methods, they can be computationally expensive, and its convergence might not be easy to diagnose [10]. In this paper we investigate an alternative approach based on the variational method [20, 19, 24]. In spite of its wide popularity in the machine learning community, the application of variational methods to inverse problems seems largely unexplored [23, 18, 17, 13]. Tipping and Lawrence (2005) [23] and Jin (2012) [17] developed Bayesian approaches to inverse problems with a heavy-tailed t model to cope data with outliers. Our proposed approach generalizes the method developed in [17] by a robust Bayesian formulation of the inverse problem using the skew-t distribution and a sparse prior structure. The attractive features of this approach are (i) uncertainty quantification of the computed solution, (ii) robustness to data outliers, and (iii) general applicability to both linear and non-linear inverse problems. We shall illustrate the efficiency of our proposed method on several ill-posed inverse problems.

The present work extends our prior work [17] in two aspects. First, this work considers the skew-t distribution for the skewness of data errors, whereas [17] considers only the t-distribution. The skewness in the error distribution introduces an extra layer of the computational complexity in developing efficient inference algorithms. Second, this work studies a sparse prior distribution, which is far more complicated than the smoothness prior analyzed in [17]. It is noteworthy that the hierarchical Bayesian model to be developed is generally applicable to linear and nonlinear inverse problems.

The rest of the paper is structured as follows. In Section 2, we formulate the inverse problem and construct the hierarchical model for our case. Then we derive the variational solution and discuss its theoretical properties in Section 3. Later, in Section 4 we illustrate the approach on two ill-posed inverse problems, i.e., the Cauchy problem and the permeability estimation in reservoir simulation, and compare its performance with the more conventional Markov chain Monte Carlo.

2. Methodology

Consider the following finite-dimensional linear inverse problem

$$\mathbf{y} = \mathbf{K}(\mathbf{u}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{K} : \mathbb{R}^m \to \mathbb{R}^n$ denotes the forward model, $\mathbf{u} \in \mathbb{R}^m$ is the unknown solution of interest, $\mathbf{K}(\mathbf{u})$ represents the model output from the forward model, and $\boldsymbol{\epsilon}$ is the additive error to the data. Thus, the vector $\mathbf{y} \in \mathbb{R}^n$ represents the noisy data that is observed or measured. Such a problem setup arises in various physical applications. One example is the Cauchy type problem for the Laplace equation, where an elliptic partial differential equation (PDE) is satisfied over a region with some over-specified boundary conditions on a part of the boundary. For example, in case of a re-entrance space shuttle, the temperature field $\mathbf{u}$ on the outer surface is to be estimated from the temperature and the flux measured at the inner surface, while an underlying PDE (steady/ transient heat equation) is satisfied. This inverse problem is severely ill-posed and a regularized solution is often sought for. In a Bayesian framework, the data is modelled statistically, and the statistical description is given by the likelihood function $p(\mathbf{y}|\mathbf{u})$, which in turn is dictated by the error distribution of the additive noise $\boldsymbol{\epsilon}$. Furthermore, we need to specify a prior distribution $p(\mathbf{u})$ on the unknown quantity $\mathbf{u}$, reflecting the prior knowledge before collecting the data. Using Bayes' theorem, we obtain the posterior distribution $p(\mathbf{u}|\mathbf{y})$ of the unknown $\mathbf{u}$

$$p(\mathbf{u}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{u})p(\mathbf{u}),$$

where $\propto$ denotes up to a multiplicative normalizing constant. This is the complete Bayesian solution of the inverse problem (1). Hence, we have to specify the likelihood function $p(\mathbf{y}|\mathbf{u})$ and the prior distribution $p(\mathbf{u})$, which constitute the two essential components of constructing the Bayesian solution. In the following two subsections, we describe the likelihood function $p(\mathbf{y}|\mathbf{u})$ and the prior distribution $p(\mathbf{u})$.

2.1. Likelihood function

In order to cope with the presence of outliers and skewness in the observational data $\mathbf{y}$, we choose to model the noisy data by a very flexible class of distributions, i.e., the skew-t distribution. The skew-t distribution, with the scale parameter, skewness parameter, and degrees of freedom, includes Gaussian, centered-t, and skew-normal distribution as special cases. It has been intensively studied since 2001, as an extension of the skew normal family, which was first introduced by Azzalini [2]. There are several different but mathematically equivalent parameterizations of skew-t distributions; see, e.g., Branco and Dey (2001, 2002) [8, 9], Azzalini and Capitanio (2003) [5]. Kim and Mallick (2003) [21] studied moments and quadratic forms of the skew-t distribution.

By assuming that each noise component ϵ_i is independent and identically distributed, following the skew-t distribution, the density function can be expressed as

$$p(\epsilon_i|\sigma^2, \alpha, \nu) \propto \frac{1}{\sigma}\left(1 + \frac{\epsilon_i^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \int_{-\infty}^{\xi_i} \left(1 + \frac{x^2}{\nu+1}\right)^{-\frac{\nu+2}{2}} dx,$$

where $\xi_i = \frac{\alpha\epsilon_i}{\sigma}\sqrt{\frac{\nu+1}{\nu+(\epsilon_i/\sigma)^2}}$, ν is the degrees of freedom, σ is a scale parameter, and α is a skewness parameter. The skew-t distribution degenerates to the regular Student's t distribution, if the skewness parameter α vanishes $\alpha = 0$. Further, when the degrees of freedom ν approaches ∞, the skew-t distribution recovers the skew normal distribution. Therefore, the normal distribution is obtained when the skewness parameter α is set to zero and the degrees of freedom ν approaches ∞. Following the preceding setup, the likelihood function $p(\mathbf{y}|\mathbf{u})$ is given by

$$p(\mathbf{y}|\mathbf{u}) \propto \left(\frac{1}{\sigma}\right)^n \prod_{i=1}^{n} \left(1 + \frac{|\mathbf{K}(\mathbf{u}) - \mathbf{y}|_i^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \int_{-\infty}^{\xi_i} \left(1 + \frac{x^2}{\nu+1}\right)^{-\frac{\nu+2}{2}} dx,$$

where

$$\xi_i = \frac{\alpha(\mathbf{K}(\mathbf{u}) - \mathbf{y})_i}{\sigma}\sqrt{\frac{\nu+1}{\nu + (|\mathbf{K}(\mathbf{u}) - \mathbf{y}|_i/\sigma)^2}}.$$

The stochastic representation of the skew-t distribution has been studied by several researchers [1, 11]. Following the representation of Cancho et al. (2011) [11], we arrive at the following hierarchical structure

$$y_i = \mathbf{K}(\mathbf{u})_i + \Delta z_i + w_i^{-\frac{1}{2}}\tau^{\frac{1}{2}}N_i,$$

where

- $z_i = w_i^{-\frac{1}{2}}|N_{0,i}|$ and $N_{0,i} \sim N(0,1)$,
- $N_i \sim N(0,1)$ and N_i is independent of $N_{0,i}$,
- $\Delta = \frac{\alpha}{\sqrt{1+\alpha^2}}\sigma$ and $\tau = \frac{\sigma^2}{1+\alpha^2}$,
- and $w_i \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$.

Likewise, the density function $f(\epsilon)$ of the skew-t distribution ξ_i can be recast as a scale mixture:

$$\begin{aligned} f(\epsilon) &= \int_0^\infty \int_0^\infty f_{\epsilon|Z,W}(\epsilon; z, w) f_{Z|W}(z; w) f(w) dz dw \\ &= \int_0^\infty \left[\int_0^\infty \frac{w\sqrt{1+\alpha^2}}{\pi\sigma} \exp\left\{-\frac{wz^2}{2} - \frac{w(1+\alpha^2)}{2\sigma^2}\left(\epsilon - \frac{\alpha\sigma z}{\sqrt{1+\alpha^2}}\right)^2\right\} dz\right] f(w) dw \\ &= \int_0^\infty \left[\frac{2\sqrt{w}}{\sigma}\phi\left(\frac{\epsilon}{\sigma/\sqrt{w}}\right)\Phi\left(\frac{\alpha\epsilon}{\sigma/\sqrt{w}}\right)\right] \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} w^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}w} dw, \end{aligned}$$

where the term $2\phi(\cdot)\Phi(\cdot)$ is the density function of the skew-normal distribution [2]. Thus, the skew-t distribution can be represented as a scale-mixture of the skew-normal distribution (more precisely a mean-scale-mixture of the normal distribution). The hierarchical representation from Cancho et al. (2011) [11] is especially suitable for posterior sampling and can also be very useful in our variational approximation of the posterior distribution.

The re-parametrization allows us to put prior on the hyperparameters Δ and τ, while maintaining the computational convenience of the model. The likelihood $p(\mathbf{y}|\mathbf{u})$ of the data $\mathbf{y}$ given $\mathbf{u}$ and the hyper-parameters $\mathbf{z}$, $\mathbf{w}$, Δ and τ is as follows

$$p(\mathbf{y}|\mathbf{u}, \mathbf{z}, \mathbf{w}, \Delta, \tau) \propto \prod_{i=1}^{n} (w_i^{-1}\tau)^{-\frac{1}{2}} \exp\left\{-\frac{w_i}{2\tau}\left(y_i - \mathbf{K}(\mathbf{u})_i - \Delta z_i\right)^2\right\}, \tag{2}$$

where $\mathbf{w}$ and $\mathbf{z}$ are the vectors of w_i's and z_i's.

2.2. Prior specification

To stably solve any ill-posed inverse problem, we have to regularize it. In the Bayesian context, the prior distribution plays the role of regularization, in the same manner the penalty function in classical regularization techniques [16]. Hence, its appropriate choice is extremely important for getting physically meaningful inverse solutions. However, the proper choice shall reflect the domain specific knowledge about concrete applications. One such example is anatomical knowledge from a complementary medical imaging modality. In this work, we shall consider a sparse type prior on the unknown $\mathbf{u}$. Our choice is motivated by following observations. First, the sparse prior represents an important prior that has successfully found numerous engineering applications, especially in the deterministic context. Meanwhile, it is illuminating for the hierarchical Bayesian modeling, since the technique can be extended other priors, e.g., the conventional smoothness prior or priors with scale-mixture representation, with little extra effort.

Specifically, we illustrate the approach with the total variation prior $p(\mathbf{u})$. It is equivalent to using a Laplace (double exponential) prior (or the Bayesian Lasso prior) [7, 22] on the differences between neighboring entries. Following the works [7, 22], we employ a scale mixture of normal representation for the Laplace distribution. Hence, the prior distribution $p(\mathbf{u})$ is given by

$$p(\mathbf{u}|\lambda) \propto \lambda^m \exp\left(-\lambda(|u_1| + |u_2 - u_1| + \cdots + |u_m - u_{m-1}|)\right). \tag{3}$$

Let $\mathbf{L}$ be an $m \times m$ matrix with $\mathbf{L}(1,1) = 1$. For $i > 1$, the ith row, denoted by $\mathbf{L}_i$, has 1 in the ith entry and -1 in the $(i-1)$th entry and the rest of the elements of the vector $\mathbf{L}_i$ are zero. Using the expression from Park and Casella (2008) [22], it can be represented by

$$p(\mathbf{u}|\lambda) \propto \int \cdots \int (\prod_{j=1}^{m} s_j)^{-1} \exp\left(-\frac{1}{2}\mathbf{u}^t\mathbf{L}^t\Sigma_s^{-1}\mathbf{L}\mathbf{u}\right) \prod_{j=1}^{m} \lambda^2 \exp\left(-\frac{\lambda^2 s_j^2}{2}\right) ds_1^2 ds_2^2 \ldots ds_m^2. \tag{4}$$

In the representation (4), Σ_s is an $m \times m$ matrix with diagonal elements s_i^2. The normal mixture representation gives a conjugate structure in the conditional posterior distribution. Let $\mathbf{s}$ be the vector of the elements s_i's. We assume a flat normal prior on the hyperparameter Δ and Gamma priors on the inverse variance (precision) τ^{-1} and the hyperparameter $\lambda_1 = \lambda^2$. In summary, we have the following scale mixture representation of the prior distribution $p(\mathbf{u}|\mathbf{s})$:

$$\begin{aligned} p(\mathbf{u}|\mathbf{s}) &\sim MVN(\mathbf{0}, (\mathbf{L}^t\Sigma_s^{-1}\mathbf{L})^{-1}), \\ p(\{s_j^2\}_{j=1}^m) &\propto \prod_{j=1}^m \lambda^2 \exp\left(-\frac{\lambda^2 s_j^2}{2}\right), \\ \Delta &\sim N(0, \sigma_\Delta^2), \\ \tau^{-1} &\sim \text{Gamma}(a_0, b_0), \\ \lambda_1 &\sim \text{Gamma}(a_1, b_1). \end{aligned} \tag{5}$$

The notation $MVN(\mathbf{m}, \Sigma)$ denotes a multivariate normal distribution with mean $\mathbf{m}$ and covariance Σ, and $\text{Gamma}(a, b)$ denotes the Gamma distribution with parameters (a, b).

By combining the representations (4) and (2) using Bayes' formula, we arrive at the following posterior distribution

$$\begin{aligned} p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \tau, \Delta, \lambda|\mathbf{y}) \propto &\prod_{i=1}^n w_i \tau^{-\frac{1}{2}} \exp\left\{-\frac{w_i z_i^2}{2} - \frac{w_i}{2\tau}\left(y_i - \mathbf{K}(\mathbf{u})_i - \Delta z_i\right)^2\right\} \\ &\times p(\mathbf{w})p(\mathbf{u}|\mathbf{s})p(\mathbf{s})p(\Delta)p(\tau)p(\lambda). \end{aligned} \tag{6}$$

The posterior distribution $p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \tau, \Delta, \lambda|\mathbf{y})$ is the full Bayesian solution to the inverse problem (1), and it encapsulates all the information about the problem. However, it is a distribution lives in potentially very high-dimensional spaces, and thus it is not directly informative about the unknown $\mathbf{u}$. Hence, it is necessary to have tools to explore the posterior state space. We shall develop an approximate inference method based on the mean field approximation for exploring the posterior state space, i.e., by constructing an approximate posterior density under the assumption of the conditional independence among different parameter components.

2.3. Variational approximation algorithm

Due to the presence of several hyper-parameters and the intractable normalizing constant, the posterior distribution (3) is not explicitly available in closed form. One way to explore the high-dimensional posterior state space is to use MCMC based methods to simulate samples from the posterior distribution. However, it is well known that the convergence of the chain is often not easy to diagnose [10], and it takes many samples to get the reliable estimates on the statistics, e.g., mean and variance. Hence, we shall take an alternative route and use the variational approximation approach. The variational method gives a fast iterative algorithm to approximate the posterior distribution, and to extract summarizing statistics, e.g., posterior mean and variance.

The idea behind variational approximations is to find a simpler and separable distribution to approximate the posterior density using Kullback-Leibler (KL) divergence, while still capturing distinct features of the posterior distribution (3) in a computationally efficient way. The KL divergence is a non-symmetric measure of the difference between two probability distributions, and it is defined as

$$\begin{aligned}
&D_{KL}(q(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau,\lambda)|p(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau,\lambda|\mathbf{y}))\\
&=\int\cdots\int q(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau,\lambda)\ln\frac{q(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau,\lambda)}{p(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\boldsymbol{\Delta},\tau,\lambda|\mathbf{y})}d\mathbf{u}d\mathbf{z}d\mathbf{w}d\mathbf{s}d\Delta d\tau d\lambda\\
&=\int\cdots\int q(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau,\lambda)\ln\frac{q(\mathbf{u},\mathbf{z},\mathbf{s},,\Delta,\tau,\lambda)}{p(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\boldsymbol{\Delta},\tau,\lambda,\mathbf{y})}d\mathbf{u}d\mathbf{z}d\mathbf{w}d\mathbf{s}d\Delta d\tau d\lambda+\ln p(\mathbf{y}),
\end{aligned}\tag{7}$$

where $p(\mathbf{y})$ is the normalizing constant of the posterior distribution (3), i.e.,

$$p(\mathbf{y})=\int\cdots\int p(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\boldsymbol{\Delta},\tau,\lambda,\mathbf{y})d\mathbf{u}d\mathbf{z}d\mathbf{w}d\mathbf{s}d\Delta d\tau d\lambda.$$

Since the term $p(\mathbf{y})$ in equation (7) is a constant, minimizing the KL distance is equivalent to minimizing the first term in equation (7). Upon slightly abusing the notation, we shall denote the first term by D_{KL} as well. In this way, we have successfully transform the sampling problem into an equivalent optimization problem of finding a simpler distribution $q(\mathbf{u},\mathbf{z},\mathbf{w},\lambda,\Delta,\tau,\lambda)$ by minimizing the KL distance D_{KL}. If we impose no constraint on the approximation $q(\mathbf{u},\mathbf{z},\mathbf{w},\lambda,\Delta,\tau,\lambda)$, minimizing the KL distance recovers the true posterior density, which however is numerically intractable. The intractability is largely due to the strong coupling between the factors. To enable the computational tractability, we impose a conditional independence condition, or equivalently separability, among the parameter components $\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau$ and λ as

$$q(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau,\lambda)=q(\mathbf{u})q(\mathbf{z})q(\mathbf{w})q(\mathbf{s})q(\Delta)q(\tau)q(\lambda).\tag{8}$$

Under this assumption, we can find an effective approximate posterior density, denoted by $q(\mathbf{u},\mathbf{z},\mathbf{w},\mathbf{s},\Delta,\tau,\lambda)$ hereafter, using an alternating direction iterative algorithm, cf. Algorithm 1 for a complete list of steps.

Remark 1. *The conditional independence assumption between different groups of the parameters decouples these factors, which then enables the computational tractability of the variational algorithm. To further reduce the computational cost, one can divide the vector parameter* $\mathbf{u}$ *into subgroups. In this work, we do not assume the conditional independence between the components of the parameter vector* $\mathbf{u}$*. The validity of such an approximation is problem dependent. This assumption represents also the essential restriction, which may compromise the accuracy of the variational approximation: Unlike the Markov chain Monte Carlo, which can be made arbitrarily accurate by running the chain sufficiently, the variational approximation has only limited accuracy. The accuracy of the variational approximation is determined by the strength of the correlation between different factors, and the weaker is the correlation, the more accurate is the approximation. However, the correlation strength generally cannot be verified a prior. To this day, the accuracy issue remains one of the open theoretical questions on the variational approximation. For related discussions of this interesting issue, we refer to the work [25].*

Next we develop the explicit formulas for carrying out each step of Algorithm 1. To this end, we first derive the necessary optimality system (with respect to each component).

Algorithm 1 Variational Bayesian approximation algorithm.

```
Set initial guess $q^0(\mathbf{w}), q^0(\mathbf{z})$, $q^0(\lambda)$, $q^0(\mathbf{s})$, $q^0(\Delta)$ and $q^0(\tau)$.
for $k = 1, \cdots, K$ do
    Find $q^k(\mathbf{u})$ by
        $q^k(\mathbf{u}) = \arg\min_{q(\mathbf{u})} D_{KL}\left(q(\mathbf{u})q^{k-1}(\mathbf{z})q^{k-1}(\mathbf{w})q^{k-1}(\lambda)q^{k-1}(\mathbf{s})q^{k-1}(\Delta)q^{k-1}(\tau)|p((\cdot),\mathbf{y})\right)$
    Find $q^k(\mathbf{z})$ by
        $q^k(\mathbf{z}) = \arg\min_{q(\mathbf{z})} D_{KL}\left(q^k(\mathbf{u})q(\mathbf{z})q^{k-1}(\mathbf{w})q^{k-1}(\lambda)q^{k-1}(\mathbf{s})q^{k-1}(\Delta)q^{k-1}(\tau)|p((\cdot),\mathbf{y})\right).$
    Find $q^k(\mathbf{w})$ by
        $q^k(\mathbf{w}) = \arg\min_{q(\mathbf{w})} D_{KL}\left(q^k(\mathbf{u})q^k(\mathbf{z})q(\mathbf{w})q^{k-1}(\lambda)q^{k-1}(\mathbf{s})q^{k-1}(\Delta)q^{k-1}(\tau)|p((\cdot),\mathbf{y})\right).$
    Find $q^k(\lambda)$ by
        $q^k(\lambda) = \arg\min_{q(\lambda)} D_{KL}\left(q^k(\mathbf{u})q^k(\mathbf{z})q^k(\mathbf{w})q(\lambda)q^{k-1}(\mathbf{s})q^{k-1}(\Delta)q^{k-1}(\tau)|p((\cdot),\mathbf{y})\right).$
    Find $q^k(\mathbf{s})$ by
        $q^k(\mathbf{s}) = \arg\min_{q(\lambda)} D_{KL}\left(q^k(\mathbf{u})q^k(\mathbf{z})q^k(\mathbf{w})q^k(\lambda)q(\mathbf{s})q^{k-1}(\Delta)q^{k-1}(\tau)p((\cdot),\mathbf{y})\right).$
    Find $q^k(\Delta)$ by
        $q^k(\Delta) = \arg\min_{q(\lambda)} D_{KL}\left(q^k(\mathbf{u})q^k(\mathbf{z})q^k(\mathbf{w})q^k(\lambda)q^k(\mathbf{s})q(\Delta)q^{k-1}(\tau)|p((\cdot),\mathbf{y})\right).$
    Find $q^k(\tau)$ by
        $q^k(\tau) = \arg\min_{q(\lambda)} D_{KL}\left(q^k(\mathbf{u})q^k(\mathbf{z})q^k(\mathbf{w})q^k(\lambda)q^k(\mathbf{s})q^K(\Delta)q(\tau)|p((\cdot),\mathbf{y})\right).$
    Check the stopping criterion.
end for
Return approximation $q(\mathbf{u})q(\mathbf{z})q(\mathbf{w})q(\lambda)q(\mathbf{s})q(\Delta)q(\tau)$.
```

In view of equation (7), the term D_{KL} is given by

$$D_{KL} = \int \cdots \int q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda) \ln \frac{q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda)}{p(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \boldsymbol{\Delta}, \tau, \lambda, \mathbf{y})} d\mathbf{u} d\mathbf{z} d\mathbf{w} d\mathbf{s} d\Delta d\tau d\lambda.$$

Then we plug the separable condition into equation (8), and enforce the normalizing conditions of all probability densities. To this end, we appeal to the associated Lagrange function $\mathcal{L}$ for the divergence D_{KL} as

$$\begin{aligned}\mathcal{L}(q(\mathbf{u}), q(\mathbf{z}), &q(\mathbf{w}), q(\lambda), q(\mathbf{s}), q(\Delta), q(\tau), \boldsymbol{\varrho}) \\ &= D_{KL} + \varrho_1 \left(\int q(\mathbf{u}) d\mathbf{u} - 1 \right) + \varrho_2 \left(\int q(\mathbf{z}) d\mathbf{z} - 1 \right) \\ &\quad + \varrho_3 \left(\int q(\mathbf{w}) d\mathbf{w} - 1 \right) + \varrho_4 \left(\int q(\lambda) d\lambda - 1 \right) + \varrho_5 \left(\int q(\mathbf{s}) d\mathbf{s} - 1 \right) \\ &\quad + \varrho_6 \left(\int q(\Delta) d\Delta - 1 \right) + \varrho_7 \left(\int q(\tau) d\tau - 1 \right),\end{aligned}$$

where the vector $\boldsymbol{\varrho} = (\varrho_1, \varrho_2, \ldots, \varrho_6, \varrho_7) \in \mathbb{R}^7$ is the vector of Lagrange multipliers (for the normalizing condition). To simplify the notation, we drop the arguments and denote the Lagrange function by $\mathcal{L}$. Following the derivation in Jin and Zou (2010) [18], we take the derivative of the Lagrange function $\mathcal{L}$ with respect to $q(\mathbf{u})$ and equate it to zero:

$$\begin{aligned}0 &= \frac{\partial}{\partial q(\mathbf{u})} D_{KL} + \rho_1 \\ &= \frac{\partial}{\partial q(\mathbf{u})} \int \cdots \int q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda) \ln \frac{q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda)}{p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \boldsymbol{\Delta}, \tau, \lambda, \mathbf{y})} d\mathbf{u} d\mathbf{z} d\mathbf{w} d\mathbf{s} d\Delta d\tau d\lambda + \rho_1 \\ &= \int \cdots \int \left(\frac{\partial}{\partial q(\mathbf{u})} \int q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda) \ln \frac{q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda)}{p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \boldsymbol{\Delta}, \tau, \lambda, \mathbf{y})} d\mathbf{u} \right) d\mathbf{z} d\mathbf{w} d\lambda d\Delta d\tau d\mathbf{s} + \rho_1. \\ &= \int \cdots \int [\ln q(\mathbf{u}) + 1 + \ln q(\mathbf{z}) + \ln q(\mathbf{w}) + \ln q(\lambda) + \ln q(\tau) + \ln q(\Delta) + \ln q(\mathbf{s}) \\ &\quad - \ln p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \Delta, \tau, \lambda, \mathbf{y})] \times q(\mathbf{z}) q(\mathbf{w}) q(\lambda) q(\tau) q(\Delta) q(\mathbf{s}) d\mathbf{z} d\mathbf{w} d\lambda d\Delta d\tau d\mathbf{s} + \rho_1.\end{aligned}$$

Upon rearranging the terms in the equation, we deduce

$$\begin{aligned}\ln q(\mathbf{u}) = - \int \cdots \int &[1 + \ln q(\mathbf{z}) + \ln q(\mathbf{w}) + \ln q(\lambda) + \ln q(\tau) + \ln q(\Delta) + \ln q(\mathbf{s}) \\ &+ \ln p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \Delta, \tau, \lambda, \mathbf{y})] \times q(\mathbf{z}) q(\mathbf{w}) q(\lambda) q(\tau) q(\Delta) q(\mathbf{s}) d\mathbf{z} d\mathbf{w} d\lambda d\Delta d\tau d\mathbf{s} + \rho_1.\end{aligned}$$

Now recall the normalizing condition $\int q(\mathbf{u}) d\mathbf{u} = 1$, and that all terms other than $p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \Delta, \tau, \lambda, \mathbf{y})$ are independent of the variable $\mathbf{u}$ and thus contribute only to the normalization condition (and also the Lagrange multiplier ϱ_1). Hence, we deduce that at a critical point $q^*(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda)$, the component $q^*(\mathbf{u})$ can be expressed as

$$\ln q^*(\mathbf{u}) = E_{q^*(\mathbf{z}) q^*(\mathbf{w}) q^*(\lambda) q^*(\tau) q^*(\Delta) q^*(\mathbf{s})} [\ln p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \boldsymbol{\Delta}, \tau, \lambda, \mathbf{y})] - \ln \mathbf{Z}_{\mathbf{q}^*(\mathbf{u})},$$

where $E_q[\cdot]$ denotes the expectation with respect to the density q. The constant term is given by

$$\ln Z_{q^*(\mathbf{u})} = \rho_1 + 1 + \int \cdots \int [\ln q^*(\mathbf{z}) + \ln q^*(\mathbf{w}) + \ln q^*(\lambda) + \ln q^*(\tau) + \ln q^*(\Delta) + \ln q^*(\mathbf{s})] d\mathbf{z} d\mathbf{w} d\lambda d\Delta d\tau d\mathbf{s}.$$

Finally, we have the general form

$$q^*(\mathbf{u}) \propto \exp\left(\int \cdots \int \ln p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \Delta, \tau, \lambda, \mathbf{y}) q^*(\mathbf{z}) q^*(\mathbf{w}) q^*(\lambda) q^*(\tau) q^*(\Delta) q^*(\mathbf{s}) d\mathbf{z} d\mathbf{w} d\lambda d\Delta d\tau d\mathbf{s} \right).$$

A similar expression holds for each of the remaining variables. Hence, we update each of the conditional posterior distribution while fixing the distributions of the remaining components. So far we have not employed the linearity of the inverse problem. For linear inverse problems, i.e., $\mathbf{K}(\mathbf{u}) = \mathbf{K}\mathbf{u}$, we can derive explicit formulas. A detailed expression of the distributions are given below under (2), (4) and (5). For $\mathbf{u}$ we derive

$$q^k(\mathbf{u}) \sim MVN((\mathbf{K}^t \Sigma_{q_{\tau,\mathbf{w}}}^{-1} \mathbf{K} + \mathbf{L}^t \Sigma_{q_s}^{-1} \mathbf{L})^{-1} \mathbf{K}^t \Sigma_{q_{\tau,\mathbf{w}}}^{-1} \mathbf{y}_\Delta, (\mathbf{K}^t \Sigma_{q_{\tau,\mathbf{w}}}^{-1} \mathbf{K} + \mathbf{L}^t \Sigma_{q_s}^{-1} \mathbf{L})^{-1}), \tag{9}$$

where $\Sigma_{q_{\tau,\mathbf{w}}}$ is an $n \times n$ diagonal matrix with the (i,i)th entry $(E_{q^*(\tau)}[\tau^{-1}] E_{q^*(w_i)}[w_i])^{-1}$. Similarly, $\Sigma_{q_\mathbf{s}}$ is an $m \times m$ diagonal entry with the (i,i)th entry $(E_{q^*(\mathbf{s})}[\frac{1}{s_i^2}])^{-1}$ and $\mathbf{y}_\Delta = \mathbf{y} - E_{q^*(\Delta)}[\Delta] E_{q^*(\mathbf{z})}[\mathbf{z}]$. For notational convenience, we drop the subscript q in the expectation $E_q[\cdot]$ in the following description.

Following the calculation from Park and Casella (2008) [22], $\frac{1}{s_j^2}$ follows an independent inverse Gaussian distribution with parameters

$$\mu = \sqrt{\frac{E(\lambda_1)}{E[(\mathbf{L}\mathbf{u})_j^2]}} \quad \text{and} \quad \lambda_2 = E[\lambda_1] = E[\lambda^2].$$

The density and the inverse Gaussian pdf f_{InG} are given by

$$\begin{aligned} q^*(\tfrac{1}{s_j^2}) &\sim \text{Inverse Gaussian}(\mu, \lambda_2), \\ f_{InG}(x, \mu, \lambda_2) &\propto x^{-\frac{3}{2}} \exp\left(-\lambda_2 \frac{(x-\mu)^2}{2\mu^2 x} \right). \end{aligned} \tag{10}$$

The latent variables z_i follow a truncated normal distribution, restricted to positive side, i.e.,

$$q^*(z_i) \sim TN(\mu_i, E[w_i^{-1}]\sigma_i^2) \tag{11}$$

with

$$\mu_i = \frac{E(\Delta)E[\tau^{-1}]}{E[\Delta^2]E[\tau^{-1}] + 1}(y_i - (\mathbf{K}E[\mathbf{u}])_i) \quad \text{and} \quad \sigma_i^2 = \frac{1}{E[\Delta^2]E[\tau^{-1}] + 1}.$$

For the latent variable Δ, we have

$$\begin{aligned} A &= E[\tau^{-1}]\sum_{i=1}^{n} E[w_i]E[z_i](y_i - (\mathbf{K}E[\mathbf{u}])_i), \\ B &= E[\tau^{-1}]\sum_{i=1}^{n} E[w_i]E_k[z_i^2] + \frac{1}{\sigma_\delta^2}, \\ q^*(\Delta) &\sim N(\tfrac{A}{B}, B^{-1}). \end{aligned} \tag{12}$$

For the variance hyperparameters $\mathbf{w}$, τ and λ_1, we have

$$\begin{aligned} w_i &\sim \text{Gamma}\left(1 + \frac{\nu}{2}, \frac{\nu}{2} + \frac{1}{2}E[\tau^{-1}(y_i - (\mathbf{K}\mathbf{u})_i - \Delta z_i)^2 + z_i^2]\right), \\ \tau^{-1} &\sim \text{Gamma}\left(a_0 + \frac{n}{2}, b_0 + \frac{1}{2}\sum_{i=1}^{n} E[w_i]E[(y_i - (\mathbf{K}\mathbf{u})_i - \Delta z_i)^2]\right), \\ \lambda_1 &\sim \text{Gamma}\left(a_1 + m, b_1 + \frac{1}{2}E[\|\mathbf{s}\|^2]\right). \end{aligned}$$

Since the updated densities are of standard form, we can compute the expectations in closed form and implement the variational algorithm in exact arithmetic.

Now we briefly comment on the computational complexity of Algorithm 1. To this end, we first observe that at each step, the updated densities for each component are of standard form and within the standard density families, and thus it amounts to updating the parameters for related densities. The density $p(\mathbf{u})$ follows a multivariate normal distribution, where the computation of the mean $\mathbf{u}^*$ involves solving a linear system, where the matrix essentially involves two matrix products. Note that in the implementation since the covariance $\Sigma_{\mathbf{u}} \equiv (\mathbf{K}^t\Sigma_{q_{\tau,\mathbf{w}}}^{-1}\mathbf{K} + \mathbf{L}^t\Sigma_{q_{\mathbf{s}}}^{-1}\mathbf{L})^{-1}$ is symmetric and positive semi-definite, it can be computed efficiently with the Cholesky decomposition. Hence, the computational complexity is dominated by computing the matrix product in the inverse covariance, computing the matrix inversion, and solving the linear system for the mean: the matrix multiplication involves $O(m^2n)$ operations (with m and n being the number of unknowns and number of data points respectively), Cholesky decomposition involves $O(m^3)$ operations, and the linear system involves $O(m^2)$ operations. The update of the vector $\mathbf{s}$, which follows an inverse Gaussian distribution componentwise, involves evaluating the term $E[\|\mathbf{L}\mathbf{u}\|^2]$, which by the bias variance decomposition is given by

$$E[\|\mathbf{L}\mathbf{u}\|^2] = \|\mathbf{L}\mathbf{u}^*\|^2 + \text{tr}(\Sigma_{\mathbf{u}}\mathbf{L}^t\mathbf{L}),$$

which again invokes computing the covariance and its complexity is of order $O(m^3)$. The same operation is required for updating the entries w_i. The cost of updating the rest of the components are cheaper, in comparison with the preceding ones. In sum, the operation complexity per iteration of the algorithm is of order $O(m^3 + m^2n)$.

So far we have focused on linear inverse problems. For nonlinear problems, we can recursively approximate the forward operator $\mathbf{K}(\mathbf{u})$ with its first-order Taylor expansion around the mean $\tilde{\mathbf{u}}$ of the current variational approximation $q^k(\mathbf{u})$, following [17]

$$\tilde{\mathbf{K}}(\mathbf{u}) = \mathbf{K}(\tilde{\mathbf{u}}) + \mathbf{J}(\mathbf{u} - \tilde{\mathbf{u}}),$$

where $\mathbf{J} = \nabla_{\mathbf{u}}\mathbf{K}(\tilde{\mathbf{u}})$ is the Jacobian of the forward map $\mathbf{K}$ with respect to $\mathbf{u}$. Iteratively applying the algorithm and correspondingly adjusting the Jacobian of the forward model $\mathbf{K}(\mathbf{u})$, we can deliver an accurate solution of the nonlinear inverse problem.

3. Theoretical properties

The variational Bayesian algorithm in Section 2 minimizes the KL distance between the joint density of the posterior distribution, with respect to the posterior distribution of each independent part of the distribution in a separable form. At the end of each iteration cycle (i.e., after updating distributions of $\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s},\ \Delta, \tau, \lambda$) we have N_0 many new parameters determining the posterior distribution of $\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}$,$\Delta, \tau, \lambda$, where N_0 is fixed. Since the distributional forms are fixed during the updating procedure, the value of the parameters uniquely determines the distribution. Let $\mathbf{\Theta}$ be the N_0 dimensional parameter vector. Then we have the following result.

Theorem 3.1. *Let $\mathbf{\Theta}_k \in \mathbb{R}^{N_0}$ be the set of parameters after kth iteration. Then $\mathbf{\Theta}_{n_k} \to \mathbf{\Theta}_0$ for some $\mathbf{\Theta}_0$ in $\mathbb{R}^{N_0}$ and some subsequence $\{n_k\}_{k\geq 1}$.*

Proof. (Sketch) It suffices to show that the sequence $\{\mathbf{\Theta}\}_k$ is bounded in $\mathbb{R}^{N_0}$. If not, some parameter (say $|\theta_i|$) would go to infinity in a subsequence. Then in that subsequence, the KL distance between $q(\cdot)$ and the joint density $p(\cdot, y)$ would go to infinity as well. However, we begin with a finite initial distance, and each step of Algorithm 1 decreases the KL distance. Thus, this scenario is infeasible and the sequence $\{\mathbf{\Theta}\}_k$ is bounded. Hence, it contains a convergent subsequence. □

Using the smoothness of the distributions and assuming the exchangability of the integral and differentiation we can conclude that the limit point $\mathbf{\Theta}_0$ is critical, i.e., the stationary condition holds. For the proof, see Appendix A.

Theorem 3.2. *For the limit point $\mathbf{\Theta}_0$, the stationary condition corresponding to the minimization of $D_{KL}(q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda)|p(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda, \mathbf{y}))$ holds.*

Remark 2. *Let $D_{KL}(\mathbf{\Theta}_k) = D_{KL}(q(\mathbf{\Theta}_k)|p(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda, \mathbf{y}))$. If a stationary point $\mathbf{\Theta}_0$ in $D_{KL}(\mathbf{\Theta})$ is the unique global optimum, then we have $\mathbf{\Theta}_k \to \mathbf{\Theta}_0$. If the sequence $\mathbf{\Theta}_k$ converges to $\mathbf{\Theta}_1$ in some other subsequence $\mathbf{\Theta}_{n_l}$, then $\mathbf{\Theta}_1$ is another stationary point with $D_{KL}(\mathbf{\Theta}_1) = D_{KL}(\mathbf{\Theta}_0)$, contradicting the uniqueness.*

A more general result is about the convergence to a local optimum, under stronger conditions. The rationale is that once the solution is in a small neighborhood around the stationary point, then it stays in that neighborhood and achieves the local minima. Suppose that $\mathbf{\Theta}_k$ converges to $\mathbf{\Theta}_0$ in a subsequence, which is a stationary point, cf. Theorem 3.2. If the Hessian of the KL distance is positive definite and continuous at $\mathbf{\Theta}_0$, the result can be summarized as follows. For the proof, we refer to Appendix B.

Theorem 3.3. *Suppose that D_{KL} has a multivariate Taylor expansion with a positive definite Hessian at $\mathbf{\Theta}_0$. Then we have $\mathbf{\Theta}_k \to \mathbf{\Theta}_0$.*

The preceding theoretical results suggest a stopping criterion for the variational algorithm. For example, one can take

$$\frac{\|\mathbf{u}_{k+1} - \mathbf{u}_k\|}{\|\mathbf{u}_k\|} < tol,$$

where *tol* is a pre-specified small tolerance, and $\mathbf{u}_k$ is the mean of the posterior approximation at iteration k. That is, the algorithm is assumed to have converged to the optimal solution, whenever the relative change of the posterior mean is within the tolerance *tol*.

4. Numerical experiments

Now we illustrate the efficiency of the proposed approach. We consider the following two examples, i.e., the Cauchy problem and the reservoir simulator example. These are exemplary of linear and nonlinear inverse problems for partial differential equations.

4.1. Cauchy problem for Laplace equation

In the Cauchy type problem, an elliptic PDE is satisfied over a specified domain $\Omega \subset \mathbb{R}^2$. The boundary Γ of the domain Ω is divided into $\Gamma = \Gamma_0 \cup \Gamma_1$, where Γ_0 and Γ_1 are accessible and inaccessible parts of the boundary, respectively. In the Cauchy problem, the temperature field and the flux are observed over the subset Γ_0. Such a scenario occurs for example in case of a re-entrant shuttle or spaceship, where the temperature on the outer surface is unknown and has to be estimated from the observations on the inner surface. This inverse problem is severely ill-posed and may not have a solution if errors are present in the observational data. Mathematically, the inverse problem for steady state heat equation can be written into

$$-\Delta u = f \tag{13}$$

with the following Cauchy type boundary conditions

$$u = g \quad \text{and} \quad \frac{\partial u}{\partial n} = q \text{ on } \Gamma_0,$$

where n is the unit outward normal direction to the boundary. The inverse problem is to estimate $\theta = u$ on the inaccessible boundary Γ_1 from the measured data g and q on a subset of Γ_0. In practical computations, we divide the domain Ω into finitely many triangles and approximate the function by the continuous piecewise linear finite element functions. In turn, the discretization gives us the following finite-dimensional formulation

$$\theta(x) = \sum_{i=1}^{m} w_j(x) u_j$$

over the boundary Γ_1 which yields the true data $\mathbf{y}_t = \mathbf{K}\mathbf{u}$, where $\mathbf{K}$ is the $n \times m$ sensitivity matrix with n the number of observations on Γ_0 and m is the number of basis elements on the boundary Γ_1.

For our numerical test, we take the domain Ω to be the unit square $\Omega = [0,1] \times [0,1]$. The boundaries Γ_1 and Γ_0 are taken to be $\Gamma_1 = (0,1) \times \{1\}$ and $\Gamma_0 = \Gamma \setminus \Gamma_1$. We take the exact solution $u(x_1, x_2)$ to the Laplace equation (13) to be $u(x_1, x_2) = \sin(\pi x_1)e^{\pi x_2} + x_1 + x_2$. We generate $y_i = y_{t,i} + \alpha_1 \max_i\{|y|_{t,i}\}e_i$. Here, (13) is satisfied with $f = 0$. The number of data points is $n = 2n_1$ on $\{0,1\} \times (0,1)$, with n_1 equally spaced points on each side of the square domain. We use m many basis on the boundary Γ_1. In our numerical simulation, we consider different values of the skewness parameter $\delta := \alpha/\sqrt{1+\alpha^2}$. The random variables e_i's follow an independent skewed-t distribution with different parameters δ and σ. Using $n_1 = 80$ and $m = 41$ we perform a simulation study for different δ values.

In this example we keep the relative noise level α_1 between 30% and 50%. Let $\hat{\mathbf{u}}$ be the posterior mean of $\mathbf{u}$ from the variational approximation. Let

$$ER1 = \frac{\|\hat{\mathbf{u}} - \mathbf{u}\|_2}{\|\mathbf{u}\|_2}$$

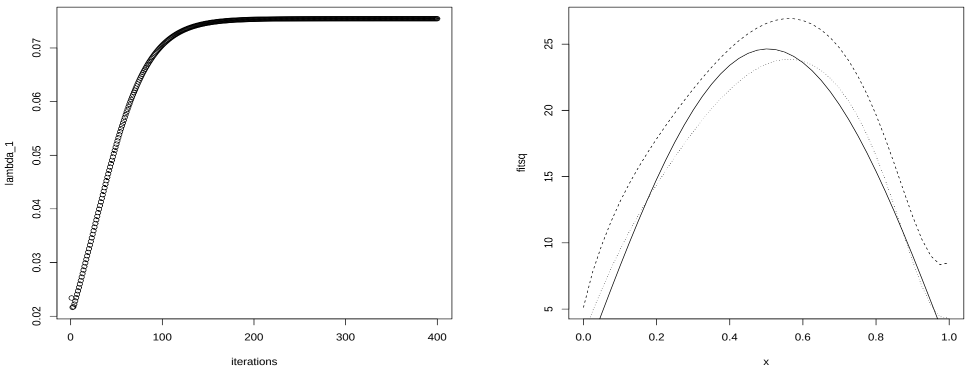
and $ER2$ be the quantity when δ is assumed to be zero. The quantity $E_f = \sqrt{ER2}/\sqrt{ER1}$ denotes the relative gain in efficiency. For the prior on the hyperparameter Δ, we use a flat normal $N(0, 100)$. The parameters for the inverse Gamma distributions are set to $(a_0, b_0) = (a_1, b_1) = (1, 1)$. We run the variational algorithm for 300 iterations in each of the cases, which gives a tolerance tol value around 10^{-6}.

Table 1: The relative efficiency and estimation of δ. The noise level is 40%, and $\nu = 4$. In the table, s_δ is standard error of the estimator $\hat{\delta}$ of δ.

δ	$\hat{\delta}$	s_δ	E_f	σ	$\hat{\sigma}$
.6	.68	.17	1.41	6.48	7.02
.8	.75	.07	2.15	6.48	6.80
.9	.86	.04	2.41	6.48	6.10

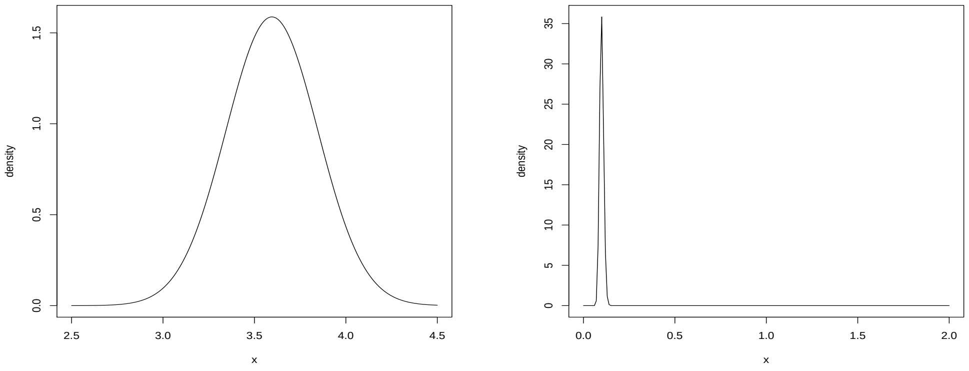
In Figure 1, we plot the recovered mean value and also the results for the more conventional approach, which ignores the skewness present in the data. We observe that accounting for the skewness of the data does give a much better recovery. The numerical results for the relative gain in efficiency due to the new model is given in Table 1. The estimated values of δ and σ are close to the true values. It is noteworthy that the convergence of the variational algorithm is almost independent of the noise level, which awaits further theoretical justification. In Figure 2, the posterior distribution of Δ and τ^{-1} are given which are normal and Gamma distribution, respectively. We note that if desired, the posterior distribution can be used to quantify the uncertainty associated with the mean, via e.g., the credible intervals, cf. Fig. 2.

The variational algorithm provides a fast, stable and accurate approximation of the posterior distribution. In contrary, the MCMC based method may take long time to converge. Even though if we run the MCMC chain long enough we can have a very accurate approximation of posterior distribution, the cost may be prohibitively high. Also, the presence of latent variables in the density may cause slow mixing and thus, makes it far more expensive computationally than the variational approach. In Table 2 a computational cost analysis is given for different parameter settings, where the proposed variational method



(a) convergence of λ_1 (b) inverse solution

Figure 1: The numerical results for the Cauchy problem for the Laplace equation, with the parameter setup $\delta = .8$, $\frac{\nu}{2} = 15$, $\alpha = 40\%$. The left panel shows the convergence of the regularizing parameter λ_1. The right panel shows a typical fit. The solid line shows the true value of the coefficient on $x = \{1\} \times [0, 1]$, the dotted line gives the recovered mean $\hat{\mathbf{u}}$ by the proposed variational algorithm, and the dashed line shows the fit with the skewness ignored.



(a) posterior density $q(\Delta)$ (b) posterior density $q(\tau^{-1})$

Figure 2: The posterior distribution for the Cauchy problem for the Laplace equation, with the parameter setup $\delta = .8$, $\frac{\nu}{2} = 15$, $\alpha = 40\%$. Left hand panel shows the posterior distribution of Δ from equation 12 and right hand panel shows the posterior Gamma distribution for τ^{-1}.

are shown to be suitable and cost efficient for high dimensional problems. Numerical experiments indicate that it is scalable with the number m of unknowns and the number n of data points. Fig. 3 shows the evolution of the posterior mean for the tuning parameter λ_1, indicating the slow convergence of the MCMC chain.

4.2. Multi-phase flow in reservoir simulation

The multiphase flow models (including water, oil, and/or gas phases) in heterogeneous porous media are widely applied in many subsurface problems, e.g., reservoir simulations. For each phase in the flow, it is governed by fluid flow equations and mass conservation. We consider the application of the variational approximation for estimating the permeability

Table 2: A comparison of the computational cost between the variation approximation and MCMC. The numerical results for the Cauchy problem for the Laplace equation, with the parameter setup $\delta = .8$, $\frac{\nu}{2} = 15$, $\alpha = 40\%$. The computational times are given in seconds. For the MCMC, the chain length is determined by the convergence of the posterior mean of λ_1.

Method	Computational cost		
	$m = 40, n_1 = 80$	$m = 60, n_1 = 120$	$m = 80, n_1 = 160$
Variational	0.82	1.96	4.20
MCMC	243.20	529.27	858

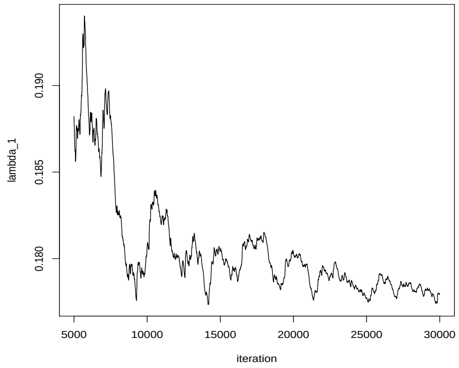


Figure 3: The posterior means of λ_1 for different chain lengths are given. The convergence is slow when compared with the variational approach in Figure 1.

field under the two-phase (water and oil) flow equations. Under the assumption that (1) fluid displacement is dominated by viscous effects, (2) gravity, compressibility, and capillary pressure are neglected, and (3) porosity is considered to be constant, we can write the governing equation of two-phase flow in terms of pressure p and saturation S as follows:

$$\begin{aligned} \nabla \cdot (\lambda(S) k \nabla p) &= Q_s, \\ \frac{\partial S}{\partial t} + v \cdot \nabla f(S) &= 0, \end{aligned}$$

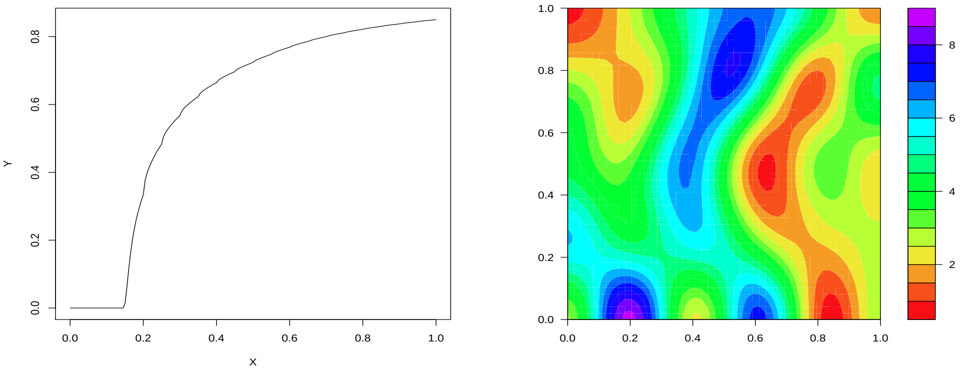
where λ is the total mobility, Q_s is the source term, k is the permeability field, v is the total velocity, and f is the fractional flux of water. Both λ and f, which are function of the saturation S, are given by the following equations,

$$\begin{aligned} \lambda(S) &= \frac{k_{rw}(S)}{\mu_w} + \frac{k_{ro}(S)}{\mu_o} \\ f(S) &= \frac{k_{rw}(S)/\mu_w}{k_{rw}(S)/\mu_w + k_{ro}(S)/\mu_o}, \end{aligned}$$

where k_{rw} is the relative permeability to water, k_{ro} is the relative permeability to oil, μ_w is the viscosity of water, and μ_o is the viscosity of oil. We use quadratic relative permeabilities, $k_{rw}(S) = S^2$ and $k_{ro} = (1 - S)^2$. The total velocity is given by the sum of phase velocities, i.e.,

$$v = v_w + v_o = -\lambda(S)k \cdot \nabla p.$$

The goal of this application is to simulate the (high-dimensional) permeability field conditioned on some observed data (e.g., pressure data and water cut data). In equation (1), $\mathbf{K}(\mathbf{u})$ denotes the integrated response, which is non-linear mapping between permeability field and the response variable. The exact water cut data and the reference permeability field are given in Figure 4. The error ϵ is independently and identically distributed as $ST(0, \sigma^2 = 1, \alpha = 0.9, \nu = 4)$. The permeability field is discretized onto a 50×50 grid. The noisy data is generated in the same way as in the Cauchy problem. However, due to the property of water cut values, which lies between 0 and 1, we introduced an additional factor to control the noise. The synthetic noisy data is generated by the following equation,



(a) true water cut data

(b) true log permeability field

Figure 4: The left hand panel shows the true data values and the right hand panel shows the true log permeability field.

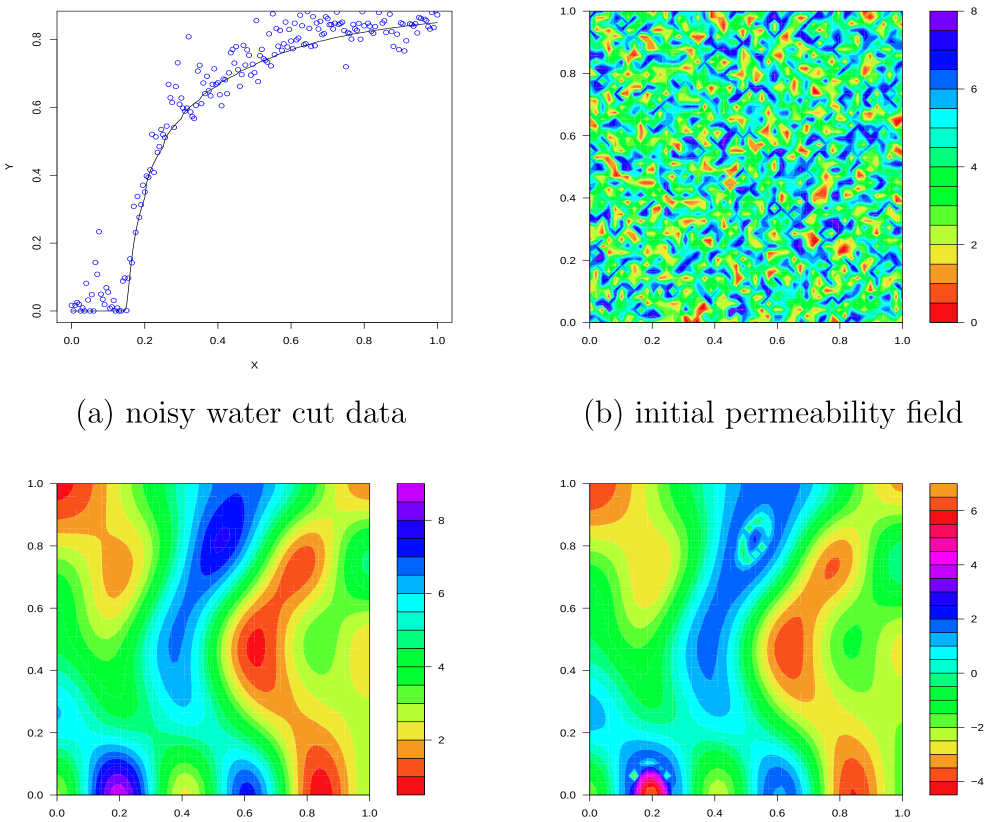
$$y_i = \begin{cases} y_i^* & \text{with prob. } r \\ y_i^* + \omega\xi\epsilon_i & \text{with prob. } 1-r, \end{cases}$$

where r is the probability of the corruption, and ω denotes the relative noise level, which controls the magnitude of the noise relatively to true response value.

The algorithm converged within 100 iterations. With higher error variance, the estimation accuracy decreases accordingly, cf. Figures 5–8. Nonetheless, even for the highest noise, the recovered permeability field remains fairly reasonable. Even though not presented, we note that the variational algorithm turns out to be much faster than the regular MCMC based algorithm.

5. Concluding Remarks

In this paper we have developed a robust Bayesian approach to inverse problems with skewed data. We use the skew-t error distribution to handle outliers and skewness in the data simultaneously. The proposed parameterization enables deriving efficient variational algorithms, as illustrated by two exemplary inverse problems. In the future, we plan to use a



(a) noisy water cut data

(b) initial permeability field

(c) log of the recovered field (mean)

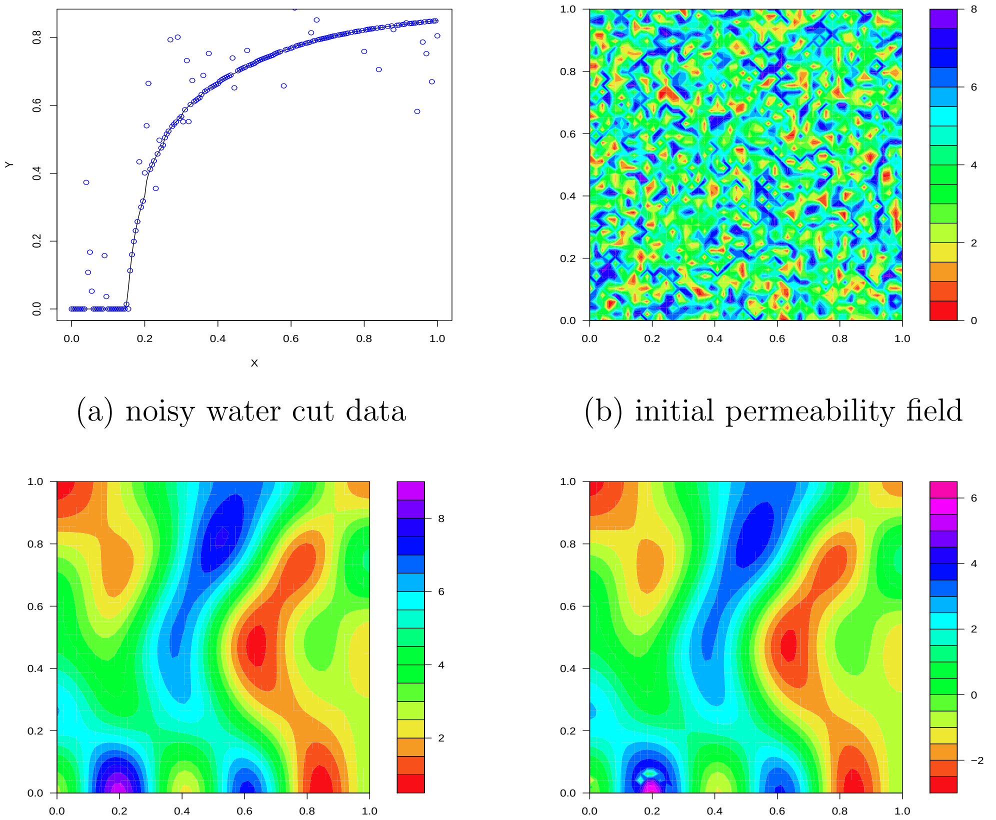
(d) log of the absolute error in the fit

Figure 5: The numerical results for the multiphase flow example, with 5% relative noise level and $r = 1$, on a 50×50 grid. Top row shows the observed data value around the true water-cut data and the initial permeability field in left and right panel, respectively. In the second row the left hand panel shows the log of the recovered field (i.e., posterior mean) and the log of the absolute error in fit given in the right hand panel.

spatially correlated prior distribution in place of the independent Lasso prior. Furthermore, we would like to study the convergence rate of the algorithm, in view of the fast convergence numerically observed in our examples. Finally, the properties, e.g., consistency, of the posterior distribution and its variational approximation, need to be established.

Acknowledgements

The authors are grateful to the anonymous referees for their constructive comments, which have led to an improved presentation of the paper. The work of B. Jin is partially supported by EPSRC grant EP/M025160/1. Y. Efendiev's work is partially supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award Number DE-FG02-13ER26165 and the DoD Army ARO Project.



(a) noisy water cut data

(b) initial permeability field

(c) log of the recovered field (mean)

(d) log of the absolute error in the fit

Figure 6: The numerical results for the multiphase flow example, with 15% relative noise level and $r = .25$, on a 50×50 grid. Top row shows the observed data value around the true water-cut data and the initial permeability field in left and right panel, respectively. In the second row the left hand panel shows the log of the recovered field (i.e., posterior mean) and the log of the absolute error in fit given in the right hand panel.

Appendix A. Proof of Theorem 3.2

We denote the value of KL distance at the kth step by D_{KL}^k. In the variational methodology, $D_{KL}^k \geq D_{KL}^{k+1} \geq 0$. Thus we have a decreasing sequence of the KL distance and thus the sequence converges to some constant c_0. The distribution of $q(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda)$ is determined by $\mathbf{\Theta}$ and is continuous in $\mathbf{\Theta}$. Thus, $D_{KL}(q(\mathbf{\Theta}_0)|p(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda, \mathbf{y})) = c_0$.

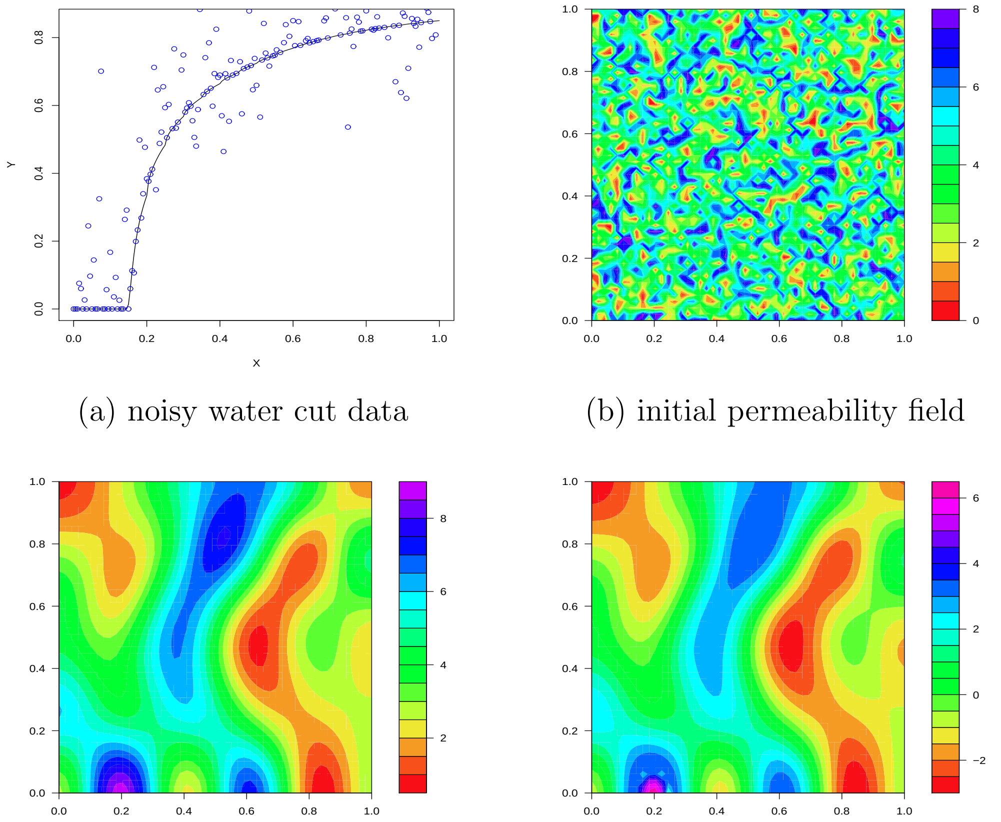
The stationary condition is satisfied if and only if

$$q^*(\mathbf{u}) \propto \exp\left(\int \cdots \int \ln p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \Delta, \tau, \lambda, \mathbf{y}) q^*(\mathbf{z}) q^*(\mathbf{w}) q^*(\lambda) q^*(\tau) q^*(\Delta) q^*(\mathbf{s}) d\mathbf{z} d\mathbf{w} d\lambda d\Delta d\tau d\mathbf{s}\right).$$

In general,

$$q^*(\eta) \propto \exp(E_{-\eta}(\ln p(\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \Delta, \tau, \lambda, \mathbf{y}))$$

where $E_{-\eta}$ denotes the expectation with respect to the proposed posterior of variables other



(a) noisy water cut data (b) initial permeability field

(c) log of the recovered field (mean) (d) log of the absolute error in the fit

Figure 7: The numerical results for the multiphase flow example, with 15% relative noise level and $r = .75$, on a 50×50 grid.Top row shows the observed data value around the true water-cut data and the initial permeability field in left and right panel, respectively. In the second row the left hand panel shows the log of the recovered field (i.e., posterior mean) and the log of the absolute error in fit given in the right hand panel.

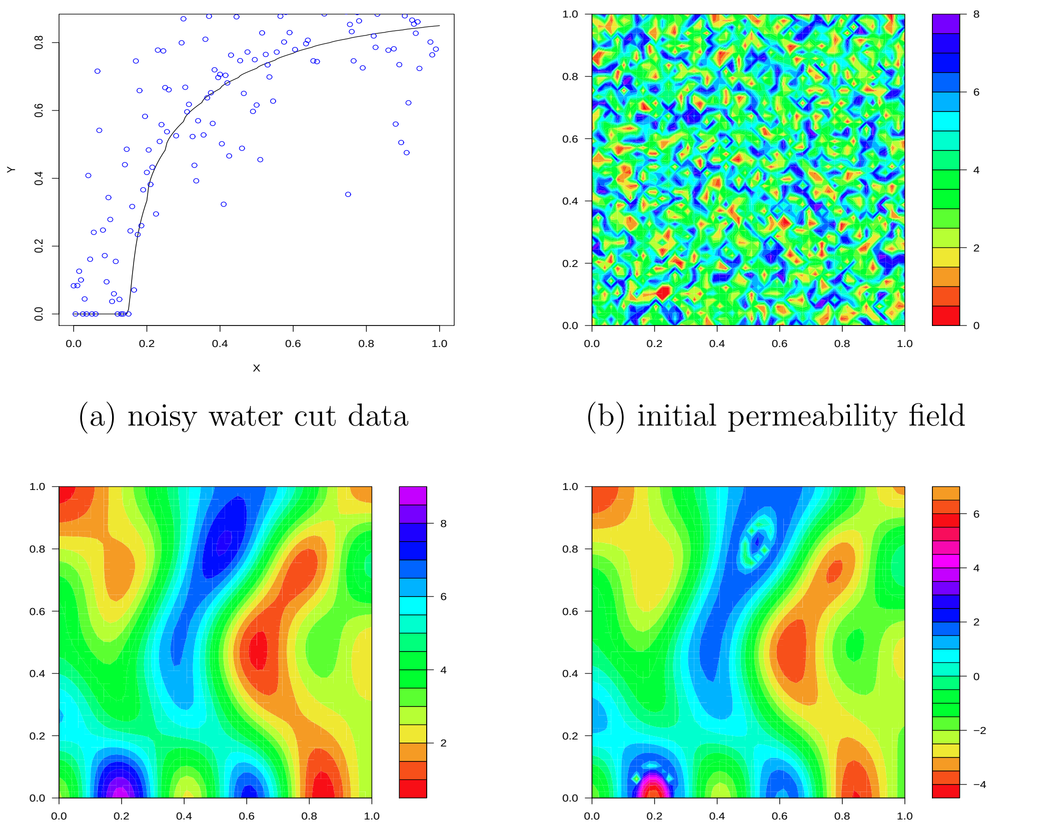
than η (for example η can be in the set $\{\mathbf{u}, \mathbf{z}, \mathbf{s}, \mathbf{w}, \Delta, \tau, \lambda\}$). If the stationary condition does not hold for some η, then further minimization of the KL distance is possible by the updating rule for η. Thus, letting $\mathbf{\Theta}_{n_k}$ to $\mathbf{\Theta}_0$, we can achieve a value $c_0 - \delta$ with $\delta > 0$. But $D_{KL}(q(\mathbf{\Theta}_0)|p(\mathbf{u}, \mathbf{z}, \mathbf{w}, \mathbf{s}, \Delta, \tau, \lambda, \mathbf{y})) = c_0$, the attained minimum value, which leads to a contradiction.

Appendix B. Proof of Theorem 3.3

(Sketch) Given $\mathbf{\Theta_0}$ there exists $\epsilon > 0$ such that

$$D_{KL}(\mathbf{\Theta}) = D_{KL}(\mathbf{\Theta}_0) + \tfrac{1}{2}(\mathbf{\Theta} - \mathbf{\Theta}_0)' H(\mathbf{\Theta}_0)(\mathbf{\Theta} - \mathbf{\Theta}_0) + o(\|\mathbf{\Theta} - \mathbf{\Theta}_0\|^2)$$

with H being the Hessian with respect to Θ and $\|\mathbf{\Theta} - \mathbf{\Theta}_0\|^2 < \epsilon$. Also, if $\|\mathbf{\Theta} - \mathbf{\Theta}_0\|^2 < \epsilon$, then $D_{KL}(\mathbf{\Theta}) - D_{KL}(\mathbf{\Theta}_0) < \epsilon_1$ implies that $\|\mathbf{\Theta} - \mathbf{\Theta}_0\|^2 < c\epsilon_1$ for some $c > 0$, because of positive definiteness of the Hessian at the stationary point $\mathbf{\Theta}_0$ and the smoothness of



(a) noisy water cut data

(b) initial permeability field

(c) log of the recovered field (mean)

(d) log of the absolute error in the fit

Figure 8: The numerical results for the multiphase flow example, with 25% relative noise level and $r = 1$, on a 50×50 grid. Top row shows the observed data value around the true water-cut data and the initial permeability field in left and right panel, respectively. In the second row the left hand panel shows the log of the recovered field (i.e., posterior mean) and the log of the absolute error in fit given in the right hand panel.

D_{KL}. Since the subsequence $\{\mathbf{\Theta}_{n_k}\}$ converges to $\mathbf{\Theta}_0$, there exists K and $\alpha < 1$ such that $\|\mathbf{\Theta}_{k+1} - \mathbf{\Theta}_0\|^2 < \epsilon$ when $\|\mathbf{\Theta}_k - \mathbf{\Theta}_0\|^2 < \alpha\epsilon$ and $\|\mathbf{\Theta}_{n_k} - \mathbf{\Theta}_0\|^2 < \alpha\epsilon$ for $k > K$. Also, $D_{KL}(\mathbf{\Theta}_{n_k}) \downarrow D_{KL}(\mathbf{\Theta}_0)$.

By choosing $\epsilon_1 > 0$ such that $c\epsilon_1 < \alpha\epsilon$, we have $D_{KL}(\mathbf{\Theta}_{n_{k+1}}) \leq D_{KL}(\mathbf{\Theta}_{n_k}) \leq \epsilon_1$. Thus, $\|\mathbf{\Theta}_{n_{k+1}} - \mathbf{\Theta}_0\|^2 < \alpha\epsilon$. Similarly, $\|\mathbf{\Theta}_{n_{k+1}} - \mathbf{\Theta}_0\|^2 < \epsilon$ and given that the KL distance D_{KL} decreases monotonically, $\mathbf{\Theta}_{n_{k+2}}$ also lies in the $\alpha\epsilon$ neighborhood and so on, which proves our claim.

References

[1] C. A. Abanto-Valle, V. H. Lachos, and D. K. Dey. Bayesian estimation of a skew-student-t stochastic volatility model. *Methodol. Comput. Appl. Prob.*, pages doi: 10.1007/s11009–013–9389–9, 2013.

[2] A. Azzalini. A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12(2):171–178, 1985.

[3] A. Azzalini. The skew-normal distribution and related multivariate families. *Scand. J. Statist.*, 32(2):159–200, 2005. With discussion by Marc G. Genton and a rejoinder by the author.

[4] A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew normal distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(3):579–602, 1999.

[5] A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 65(2):367–389, 2003.

[6] A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

[7] K. Bae and B. K. Mallick. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–3430, 2004.

[8] M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.*, 79(1):99–113, 2001.

[9] M. D. Branco and D. K. Dey. Regression model under skew elliptical error distribution. *J. Math. Sci. (N.S.) (Delhi)*, 1:151–168, 2002.

[10] S. P. Brooks. Discussion on the Paper by Spiegelhalter, Best, Carlin, and van der Linde. *J. R. Stat. Soc. Ser. B*, 64(4):616–618.

[11] V. G. Cancho, D. K. Dey, V. H. Lachos, and M. G. Andrade. Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: estimation and case influence diagnostics. *Comput. Statist. Data Anal.*, 55(1):588–602, 2011.

[12] N. Chan, M. Getmansky, S. M. Haas, and A. W. Lo. Systemic risk and hedge funds. In M. Carey and R. M. Stulz, editors, *The Risks of Financial Institutions.* University of Chicago Press, Chicago, 2006.

[13] M. Gehre and B. Jin. Expectation propagation for nonlinear inverse problems - with an application to electrical impedance tomography. *J. Comput. Phys.*, 259:513–535, 2014.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

[15] M. G. Genton. Skew-symmetric and generalized skew-elliptical distributions. In *Skew-Elliptical Distributions and their Applications*, pages 81–100. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[16] K. Ito and B. Jin. *Inverse Problems: Tikhonov Theory and Algorithms*, volume 22 of *Series on Applied Mathematics*. World Scientific, Hackensack, NJ, 2015.

[17] B. Jin. A variational Bayesian method to inverse problems with impulsive noise. *J. Comput. Phys.*, 231(2):423–435, 2012.

[18] B. Jin and J. Zou. Hierarchical Bayesian inference for ill-posed problems via variational method. *J. Comput. Phys.*, 229(19):7317–7343, 2010.

[19] M. I. Jordan. Graphical models. *Statist. Sci.*, 19(1):140–155, 2004.

[20] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.

[21] H.-M. Kim and B. K. Mallick. A note on Bayesian spatial prediction using the elliptical distribution. *Statist. Probab. Lett.*, 64(3):271–276, 2003.

[22] T. Park and G. Casella. The Bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686, 2008.

[23] M. E. Tipping and N. D. Lawrence. Variational inference for student-t models: robust Bayesian interpolation and generalised component analysis. *Neurocomput.*, 69(1–3):123–141, 2005.

[24] D. M. Titterington. Bayesian methods for neural networks and related models. *Statist. Sci.*, 19(1):128–139, 2004.

[25] B. Wang and D. M. Titterington. Inadequacy of interval estimates corresponding to variational bayesian approximations. In R. Cowell and Z. Ghahramani, editors, *Proc. 10th Intern. Workshop Artif. Intell. Stat.*, pages 373–380. 2005.