Advances in Applied Mathematics 应用数学进展, 2024, 13(1), 349-359
Published Online January 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.131037

Hans汉斯

基于贝叶斯推断的地下水流模型非高斯源项场的识别

黄千叶[1]，欧　娜[1*]，宋晓燕[2]

[1]长沙理工大学数学与统计学院，湖南　长沙
[2]湖南工商大学理学院，湖南　长沙

收稿日期：2023年12月25日；录用日期：2024年1月19日；发布日期：2024年1月25日

摘　要

地下水模型中非均匀介质的复杂性及观测数据的稀缺性使模型存在不确定性。为了更好地预测模型的输出，需要我们基于有限的观测数据来估计模型中的未知参数，降低不确定性。贝叶斯方法是刻画不完全数据、模型偏差和测量误差带来的不确定性的有效方式，它可以根据现有数据确定参数向量的后验分布。在实际应用中，其主要挑战在于从后验分布中抽样。传统抽样方法随未知参数维数的增加而出现退化现象。本文主要利用最大期望变量选择(EMVS)方法来识别稀疏离散余弦变换(DCT)系数，并对后验分布中的超参数进行自适应更新，以提高问题的求解效率；特别地，利用逆Hessian矩阵来加速朗之万动力学蒙特卡洛马尔科夫链(MCMC)的收敛速度，使用敏感性矩阵构造的简化模型来有效地计算梯度和Hessian矩阵，高效地解决高维不确定性分析和反演问题。基于地下水源项识别高维反演数值实验，验证了反演方法能够得到可靠的参数估计，为提高地下水模拟在实际应用中的可靠性和计算效率提供了新的思路，对后续的地下水资源管理决策制定具有重要意义。

关键词

最大期望变量选择，非高斯随机场，Langevin蒙特卡洛马尔科夫链

Research on Identification of Non-Gaussian Source Term Field in Subsurface Flow Model Based on Bayesian Inference

Qianye Huang[1], Na Ou[1*], Xiaoyan Song[2]

[1]School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha Hunan
[2]School of Science, Hunan University of Technology and Business, Changsha Hunan

[*]通讯作者。

文章引用: 黄千叶, 欧娜, 宋晓燕. 基于贝叶斯推断的地下水流模型非高斯源项场的识别[J]. 应用数学进展, 2024, 13(1): 349-359. DOI: 10.12677/aam.2024.131037

Received: Dec. 25th, 2023; accepted: Jan. 19th, 2024; published: Jan. 25th, 2024

Abstract

Groundwater model of heterogeneous media and its scarcity of observation data and other related factors makes the existence of uncertainty. In order to better predict the output of the model, we need to estimate the input and parameters of the model based on limited observational data to reduce the uncertainty. Bayesian method is an effective way to describe the uncertainty caused by incomplete data, model deviation and measurement error. It can determine the posterior distribution of parameter vectors according to the existing data. In practical application, the main challenge lies in sampling. The traditional sampling method degrades with the increase of the dimension of unknown parameters, that is, the convergence is slow. In this paper, an inversion algorithm is proposed to identify sparse discrete cosine transform (DCT) coefficients in expectation maximized variable selection (EMVS) frame, and adaptively update the hyperparameters to improve the solving efficiency of the problem. In particular, the inverse Hessian is used to accelerate the convergence of Langevin dynamics Monte Carlo Markov Chain (MCMC), and the simplified model is used to compute the gradient and Hessian effectively to solve the high dimensional uncertainty analysis and inversion problems efficiently. Based on the high-dimensional inversion numerical experiment of groundwater source item identification, it is verified that the inversion method can obtain reliable parameter estimation, which provides a new idea for improving the reliability and computational efficiency of groundwater simulation in practical application, and has important significance for the subsequent decision-making of groundwater resource management.

Keywords

EMVS, Non-Gaussian Random Field, Langevin MCMC

1. 引言

地下流体流动预测模型是地下水和油气资源开发，以及场地修复和清理活动的开发和管理的重要一环。当有可靠的模型输入时，基于地下水模型的石油生产和管理是十分有效的。然而，在实践中，地下岩石层的复杂性、地区差异性和不可接近性，给储层油气勘探和开采带来了很大困难。人们通过一定的观测数据，对模型中的未知参数进行估计，进而做出合理的预测用于指导生产。

贝叶斯方法[1] [2]是参数估计的一个有效方法，它可以将观测数据中的不确定性和未知参数的先验信息相结合。相比于传统的优化方法，贝叶斯方法给出了参数的后验概率密度而不仅仅是单个的点估计，有利于我们量化参数中的不确定性。然而，正演模型的非线性性或先验信息的非高斯性，使得我们无法获得后验概率密度函数的显式表达式。

蒙特卡洛马尔科夫链(MCMC) [3] [4]方法通过构造一条或多条以后验概率密度为目标分布的马尔科夫链，可以有效地刻画后验概率分布。由于所谓的维数灾难问题，MCMC 算法在高维模型空间中的采样能力严重下降[5]。在过去的几十年里，人们提出了许多 MCMC 方法来缓解这个问题。一系列与参数维数无关的抽样方法[6]被提了出来，其中预选概率密度由离散随机微分方程(SDE)构建。当预选概率密度

推广到 Langevin 动力方程的框架下，可以得到 Metropolis 调整的 Langevin 算法(MALA) [7]，这属于 Langevin 蒙特卡洛(LMC)方法的范畴。这类方法的另一个分支为无调整的 Langevin 算法(ULA) [7]，即样本的更新由 Langevin 动力方程的直接离散获得。由于略过了 Metropolis-Hasting 方法中是否接受这一步，ULA 方法在抽样时是非常高效的。

另一方面，采用适当的方法对未知随机场参数化，降低未知参数的维数，可以降低抽样方法的计算量。一种高效且直接的方法是使用标准正交线性变换。在流行的压缩基中，Karhunen-Loève 或主成分分析最广泛地用于基于先验协方差(二阶)信息的参数化和模型约简[8] [9] [10]。Sarma 等人[11]提出了一种用于参数化非高斯通道的非线性方法来保持高阶统计量。Jafarpour 和 McLaughlin [12]引入了一种稳定且计算效率高的参数化方法，离散余弦变换 DCT，用于参数化。DCT 方法的突出的优势是其构造不需要基于未知参数的先验信息，这使它区别于 KLT 等数据依赖型参数化方法。Sahni 和 Horne [13]将另一种基于变换的参数化方法，离散小波变换(DWT)，用于历史匹配。

为了合理地估计 DCT 基函数的稀疏系数和正则化参数，本文采用了贝叶斯方法。我们使用期望最大化变量选择(EMVS)方法[14]来识别未知域的稀疏表示。采用“spike-and-slab”高斯混合先验来描述 DCT 基的稀疏性。不同的方差参数被分配给“spike”分布和“slab”分布。潜在变量的期望等价于在两个先验中选择正则化参数，它避免了正则化参数的人工选择，因为这些参数是由观测数据自然决定的。在期望最大化(EM)方法的最大化步骤中，我们使用 Langevin 动力学 MCMC 方法对 DCT 基函数的系数进行采样，在动态系统中注入的噪声有助于粒子从局部模式中逃逸。此外，我们使用目标函数的逆 Hessian 作为预处理，以加速马尔可夫链的收敛。

本文的大纲如下，在第二章中，我们着重描述了在 EMVS 框架下识别 DCT 系数的具体过程。在第三章中，我们将该方法应用于地下水流模型非高斯源项场的识别，探讨和比较了该方法在不同预处理矩阵下的收敛效果，并在参数样本的基础上对模型输出变量做出了预测。

2. 贝叶斯稀疏识别方法

2.1. 随机场的参数化

通常，信号的大部分能量都用低阶 DCT 系数表示。因此，这种数学变换可以用于模型压缩，这是通过将基函数项的系数设置为超过某个阈值等于零(截断)来实现的。在这种情况下，与保留的基相关联的系数成为表达信号的参数，而在反问题中，它们是要检索的未知参数。此外，由于其近似能力、数据无关(预构造)基础和计算复杂度低，离散余弦变换已经成功并广泛地应用于信号及图像处理领域。长度为 N 的离散一维 DCT [12]的形式如下：

$$v(k)=\alpha(k)\sum_{n=1}^{N}\cos\left[\frac{\pi(2n-1)(k-1)}{2N}\right],1\le k\le N,$$

其中 $\alpha(k)$ 被定义为

$$\alpha(k=1:N)=\sqrt{2}\alpha(0)=\sqrt{\frac{2}{N}}.$$

在多维空间中，DCT 基由一维基的元素之间的笛卡尔积得到，这允许有效的、可分离的处理多维信号。设一个二维储层离散参数场，在 x 方向有 M 个网格，y 方向有 N 个网格。对该参数场进行离散余弦变换为

$$\Phi(m,n)=\alpha_M(m)\alpha_N(n)\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\cos\left[\frac{\pi(2i+1)m}{2M}\right]\cos\left[\frac{\pi(2j+1)n}{2N}\right],$$

其中 $m=1,\cdots,M$ ， $n=1,\cdots,N$ ，且

$$\alpha_M(m)=\begin{cases}\sqrt{\dfrac{1}{M}}, & m=0\\ \sqrt{\dfrac{2}{M}}, & m\neq 0\end{cases},\quad \alpha_N(n)=\begin{cases}\sqrt{\dfrac{1}{N}}, & n=0\\ \sqrt{\dfrac{2}{N}}, & n\neq 0\end{cases}.$$

记一维 DCT 基函数矩阵为 $I_x\in\mathbb{R}^{M\times M}$ ，其元素为

$$[I_x]_{i,m}=\alpha(i)\cos\left[\frac{\pi(2m-1)(i-1)}{2M}\right],$$

且 $I_y\in\mathbb{R}^{N\times N}$ ，其元素为

$$[I_y]_{j,n}=\alpha(j)\cos\left[\frac{\pi(2n-1)(j-1)}{2N}\right],$$

则二维 DCT 基函数可以写为如下矩阵形式

$$\Phi=I_x\otimes I_y,$$

其中 $\otimes$ 表示张量积。

2.2. EMVS

记 $\mathbf{d}\in\mathbb{R}^{n_d}$ 为观测数据， $\boldsymbol{\xi}\in\mathbb{R}^p$ 是待估计的未知参数。我们假设输入的 $\boldsymbol{\xi}$ 和数据之间的关系为

$$\mathbf{d}=\mathbf{G}(\Phi\boldsymbol{\xi})+\varepsilon,$$

其中 $\Phi\in\mathbb{R}^{n\times p}$ 是正交矩阵，其列是一组正交压缩基，本文采用离散余弦变化基函数作为压缩基。 $\mathbf{G}:\boldsymbol{\xi}\to\mathbb{R}^{n_d}$ 为正演模型， $\varepsilon\sim\mathcal{N}(0,\sigma^2 I)$ 是独立同分布的噪声。那么似然函数可以写成

$$\pi(\mathbf{d}\,|\,\boldsymbol{\xi})=\left(2\pi\sigma^2\right)^{-\frac{n_d}{2}}\exp\left(-\frac{\left\|\mathbf{d}-\mathbf{G}(\Phi\boldsymbol{\xi})\right\|_2^2}{2\sigma^2}\right).$$

根据 EMVS 方法[14]，引入一个与回归系数 $\boldsymbol{\xi}$ 相同维数的二元潜在向量 γ 来表示每个系数 $\boldsymbol{\xi}_j$ 是“入还是出”。假设二进制向量 γ 上的先验分布为模型空间上的先验分布。给定 $\gamma_j=0$ 的 ξ_i 的先验通常是集中在 0 附近的正态分布，称为“ spike”先验；给定 $\gamma_j=1$ 的 ξ_i 的先验是平板或扩散分布，称为“slab”先验。利用潜在向量 γ 的后验分布来识别后验概率最高的模型。我们假设参数 ξ_i 遵循“spike-and-slab”高斯混合先验

$$\pi(\xi_i\,|\,\sigma^2,\gamma_i)=(1-\gamma_i)N(0,\sigma^2 v_0)+\gamma_i N(0,\sigma^2 v_1),$$

其中 N 为正态分布， v_0 是“spike”参数， v_1 是“slab”参数，且 $0\le v_1<v_0$ 。在参数化目标函数后的 DCT 域中，尽管有大量的基函数，但通常只有少数基函数与目标函数有真正的联系。因此，我们有理由假设真正的 $\boldsymbol{\xi}$ 是稀疏的，即使 p 在增长，其非零元素始终是有限数量的，我们的目标是识别非零元素。

误差方差 σ^2 遵循逆伽马先验 $\pi(\sigma^2\,|\,\gamma_i)=IG(\nu/2,\nu\lambda/2)$ ，恰当的先验分布可以使得后验分布的显式表达形式。当参数 $\boldsymbol{\xi}_i$ 较小或较大时，引入的二进制潜在变量 γ_i 等于 0 或 1。另外， γ 的先验遵循伯努利分布， $\pi(\gamma\,|\,\theta)=\theta^{|\gamma|}(1-\theta)^{p-|\gamma|}$ ，它包含了关于模型中需要包含哪些 ξ_i 的不确定性。这里， $|\gamma|=\sum_j^p\gamma_j$ 以及 θ 服从 $\pi(\theta)=\theta^{a-1}(1-\theta)^{b-1}$ ，其中 $a=b=1$ 产生唯一的超先验 $\theta\sim U(0,1)$ 。

由此，后验分布有如下形式

$$\pi\left(\xi,\sigma^2,\theta,\gamma\,|\,\mathbf{d}\right)\propto\pi\left(\mathbf{d}\,|\,\xi,\sigma^2\right)\pi\left(\xi\,|\,\sigma^2,\gamma\right)\pi\left(\sigma^2\,|\,\gamma\right)\pi\left(\gamma\,|\,\theta\right)\pi\left(\theta\right).$$

潜在的包含指标 γ 被视为“缺失数据”。由于这个函数是不可观察的，它在每次迭代中都被给定观察数据和当前参数估计的条件期望所取代，即所谓的 E 步。在第 k 次迭代中，EM 算法通过迭代最大化下面的目标函数来间接最大化 $\pi\left(\xi,\sigma^2,\theta\,|\,\mathbf{d}\right)$，

$$Q\left(\xi,\sigma,\theta\,|\,\xi_k,\sigma_k,\theta_k\right)=\mathbb{E}_{\gamma|\cdot}\left[\log\pi\left(\xi,\sigma,\theta,\gamma\,|\,\mathbf{d}\right)|\,\xi_k,\sigma_k,\theta_k,\mathbf{d}\right],$$

其中 $\mathbb{E}_{\gamma|\cdot}\left(\cdot\right)$ 表示条件期望 $\mathbb{E}_{\gamma|\xi_k,\sigma_k,\theta_k,\mathbf{d}}\left(\cdot\right)$。这一步骤也就是我们算法中期望的 E 步计算。对于上述共轭“spike-and-slab”混合分层先验公式，目标函数可以分离为

$$Q\left(\xi,\sigma,\theta\,|\,\xi_k,\sigma_k,\theta_k\right)=C+Q_1\left(\xi,\sigma\,|\,\xi_k,\sigma_k,\theta_k\right)+Q_2\left(\theta\,|\,\xi_k,\sigma_k,\theta_k\right),$$

其中

$$\begin{aligned}Q_1\left(\xi,\sigma\,|\,\xi_k,\sigma_k,\theta_k\right)=&-\frac{\left\|\mathbf{d}-\mathbf{G}\left(\Phi\xi\right)\right\|_2^2}{2\sigma^2}-\frac{n_d-1+p+\nu}{2}\log\left(\sigma^2\right)\\&-\frac{\nu\lambda}{2\sigma^2}-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\xi_i^2\mathbb{E}_{\gamma|\cdot}\left[\frac{1}{v_0\left(1-\gamma_i\right)+v_1\gamma_i}\right],\end{aligned}$$

$$Q_2\left(\theta\,|\,\xi_k,\sigma_k,\theta_k\right)=\sum_{i=1}^{p}\log\left(\frac{\theta}{1-\theta}\right)\mathbb{E}_{\gamma|\cdot}\gamma_i+\left(a-1\right)\log\left(\theta\right)+\left(p+b-1\right)\log\left(1-\theta\right).$$

如在[14]中所讨论的，$\mathbb{E}_{\gamma|\cdot}\left[\gamma_i\right]$ 可以计算为

$$\mathbb{E}_{\gamma|\cdot}\left[\gamma_i\right]=\mathrm{P}\left(\gamma_i=1\,|\,\xi_k,\sigma_k,\theta_k\right)=q_i,$$

其中

$$q_i=\frac{a_i}{a_i+b_i},$$

$$\begin{aligned}a_i&=\pi\left(\xi_k\,|\,\sigma_k,\gamma_i=1\right)\mathrm{P}\left(\gamma_i=1\,|\,\theta_k\right)=\pi\left(\xi_k\,|\,\gamma_i=1\right)\theta_k,\\b_i&=\pi\left(\xi_k\,|\,\sigma_k,\gamma_i=0\right)\mathrm{P}\left(\gamma_i=0\,|\,\theta_k\right)=\pi\left(\xi_k\,|\,\gamma_i=0\right)\left(1-\theta_k\right).\end{aligned}$$

函数 Q_1 中的期望可以计算为

$$\mathbb{E}_{\gamma|\cdot}\left[\frac{1}{v_0\left(1-\gamma_i\right)+v_1\gamma_i}\right]=\frac{1-q_i}{v_0}+\frac{q_i}{v_1}=d_i.\tag{1}$$

在 M 步中，我们分别最大化目标函数 Q_1 和 Q_2，来分别更新 $\left(\xi,\sigma\right)$ 和超参数 θ。通过在这两个步骤之间迭代，EM 算法生成了一系列参数估计。下降搜索方案通常需要前向模型的梯度信息，并会陷入非线性非凸目标函数[15]的局部最小值。

2.3. Langevin 动力学 MCMC

如上一节 EMVS 方法所介绍的，ξ 的优化目标函数 $Q_1\left(\xi,\sigma\,|\,\xi_k,\sigma_k,\theta_k\right)$，通过最大化目标函数可得到 ξ_{k+1} 的值，即

$$\begin{aligned}\xi_{k+1}&=\arg\max_{\xi\in\mathbb{R}^p}Q_1\left(\xi,\sigma\,|\,\xi_k,\sigma_k,\theta_k\right)\\&=\arg\min\left\{\left\|\mathbf{d}-\mathbf{G}\left(\Phi\xi\right)\right\|^2+\left\|\mathbf{D}^{1/2}\xi\right\|^2\right\},\end{aligned}$$

其中 $\mathbf{D}\in\mathbb{R}^{p\times p}$ 为一个对角矩阵 $\mathrm{diag}\{d_i\}$，其对角元素由方程(1)计算。我们借鉴了论文[16] [17]的思想，利用朗之万动力学 MCMC 方法更新目标参数 $\boldsymbol{\xi}$。动力学被定义为

$$\mathrm{d}\boldsymbol{\xi}_t=\Sigma\nabla_{\boldsymbol{\xi}}Q_1\left(\boldsymbol{\xi}_t\mid\boldsymbol{\xi}_k,\sigma_k,\theta_k\right)\mathrm{d}t+\sqrt{2\Sigma}\mathrm{d}\mathbf{W}_t,\tag{2}$$

其中 Σ 是一个任意的对称正定矩阵，$\mathbf{W}_t\in\mathbb{R}^n$ 是一个标准的 n 维布朗运动，它可以帮助粒子 $\boldsymbol{\xi}$ 逃离局部域，探索全局域。我们使用 Euler-Maruyama 格式[18]来离散方程(2)，构造马尔可夫链

$$\boldsymbol{\xi}_{k+1}=\boldsymbol{\xi}_k+\delta\Sigma\nabla_{\boldsymbol{\xi}}Q_1\left(\boldsymbol{\xi}_k\mid\boldsymbol{\xi}_k,\sigma_k,\theta_k\right)+\sqrt{2\delta\Sigma}\mathbf{w}_k,$$

其中，δ 是时间步长，$\mathbf{w}_k\sim N(0,\mathbf{I})$。预处理矩阵 Σ 被设置为 $\mathbf{I}$ 或者在 $\boldsymbol{\xi}_k$ 点处的局部逆黑森[19] $\mathbf{H}^{-1}$，以指导采样过程中的更新方向，其中

$$\mathbf{H}(\boldsymbol{\xi})=\nabla_{\boldsymbol{\xi}}^2Q_1\left(\boldsymbol{\xi}\mid\boldsymbol{\xi}_k,\sigma_k,\theta_k\right).$$

超参数 σ 和 θ 分别通过最大化 Q_1 和 Q_2 来更新。综上所述，算法 1 列出了未知数后验的采样。我们在后面的数值例子中讨论了预处理矩阵 Σ 的影响。

Algorithm 1. EMVS 方法下的 DCT 系数识别

Input：初始化 $\boldsymbol{\xi}$，σ，θ，给定步长 δ

1：**for all** $k\leftarrow 1$：

2：$a_i\leftarrow\pi\left(\boldsymbol{\xi}_k\mid\sigma_k,\gamma_i=1\right)\theta_k$，$b_i\leftarrow\pi\left(\boldsymbol{\xi}_k\mid\sigma_k,\gamma_i=0\right)\left(1-\theta_k\right)$

3：$q_i\leftarrow\dfrac{a_i}{a_i+b_i}$

4：$d_i\leftarrow\dfrac{1-q_i}{v_0}+\dfrac{q_i}{v_1}$

5：$\boldsymbol{\xi}_{k+1}\leftarrow\boldsymbol{\xi}_k+\delta\nabla_{\boldsymbol{\xi}}Q_1\left(\cdot\mid\mathbf{d}\right)+N\left(0,2\delta\right)$

6：$\sigma_{k+1}\leftarrow\sqrt{\dfrac{\left\|\mathbf{d}-\mathbf{G}\left(\Phi\boldsymbol{\xi}_{k+1}\right)\right\|_2^2+\left\|\mathbf{D}^{1/2}\boldsymbol{\xi}_{k+1}\right\|_2^2+\nu\lambda}{n-1+p+\nu}}$

7：$\theta_{k+1}\leftarrow\dfrac{\sum_{i=1}^{p}q_i+a-1}{a+b+p-2}$

3. 数值算例

在本节中，我们使用所提出的方法来反演识别未知源项场。我们考虑地下水流模型，它由下面的抛物线方程来描述，

$$\frac{\partial u(x,t)}{\partial t}=\mathrm{div}\left(\kappa(x)\nabla u(x,t)\right)+f(x),\quad x\in\Omega,\ t\in(0,T],\tag{3}$$

其中模型求解区域为 $[0,1]\times[0,1]$，$T=0.1$。我们的目标是反演未知源函数 $f(x)$，我们将边界条件设为齐次 Dirichlet 边界，初始条件设为 $u_0(x)=0$，$\kappa(x)=1$。观测数据为 T 时刻区域边界上的通量，即 $\left.\dfrac{\partial u}{\partial\mathbf{n}}\right|_{\partial\Omega}$，其中 $\mathbf{n}$ 为单位外法向量。

为了进行模拟，我们需要在有限维空间中表示函数 $f(x)$。因此我们沿用前文介绍的离散余弦变化再参数化技术，将源函数投影到由 p 个 DCT 基函数跨越的空间上。即在 DCT 域上，源函数可以表示为

$$f\left(x_j;\xi\right)=\sum_{i=1}^{p}\xi_i\Phi\left(j,i\right),$$

其中ξ是待反演的未知参数，$\Phi\in\mathbb{R}^{n\times p}$是截断的 DCT 基函数。由于方程(3)的解线性依赖于源函数，我们有以下近似：

$$\mathbf{G}\left(\Phi\xi\right)=\mathbf{S}\xi$$

其中$\mathbf{S}$为敏感性矩阵，由以下表达式定义：

$$\mathbf{S}=\left[\left.\frac{\partial u\left(\Phi_1\right)}{\partial\mathbf{n}}\right|_{\partial\Omega},\left.\frac{\partial u\left(\Phi_2\right)}{\partial\mathbf{n}}\right|_{\partial\Omega},\cdots,\left.\frac{\partial u\left(\Phi_p\right)}{\partial\mathbf{n}}\right|_{\partial\Omega}\right]$$

在这里$u\left(\Phi_i\right)\in\mathbb{R}^{n_d}$表示右端源项为$\Phi_i$时，正演模型的相关输出变量，$\Phi_i$为矩阵$\Phi$的第$i$列。

观测数据由真实的$f\left(x\right)$产生，真实的$f\left(x\right)$如图 1 所示。在生成观测数据时，我们采用的物理网格为 45 × 45，时间上的离散步长为$\Delta t=0.001$。为了避免出现“反问题陷阱”(inverse crime)这一现象在反演过程中求解正问题所采用的时间步长为$\Delta t=0.002$。我们将应用前文提到的贝叶斯方法来反演未知的随机场。

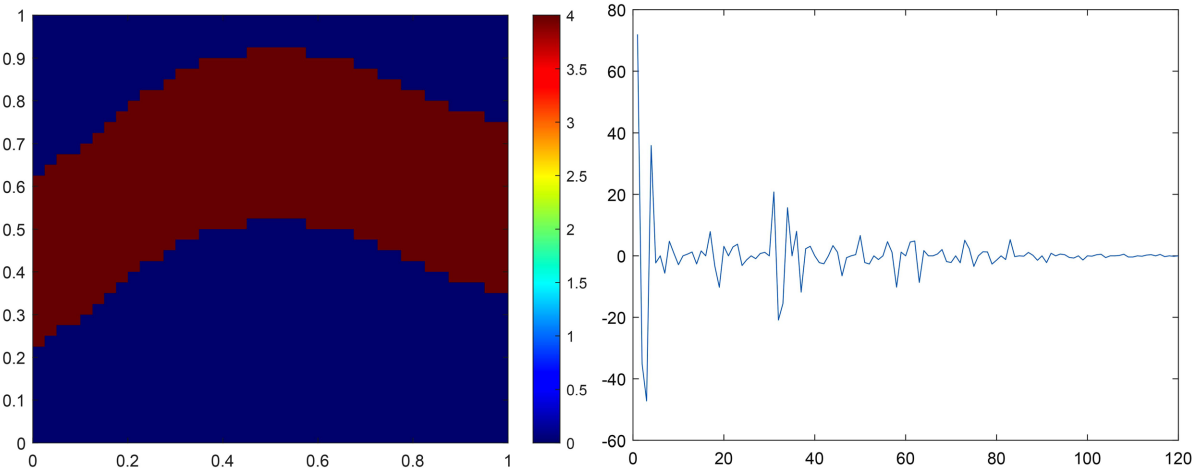


Figure 1. (a) The spatial distribution of the true $f\left(x\right)$ and (b) the DCT coefficients of the true $f\left(x\right)$

图 1. (a) $f\left(x\right)$ 的真实分布情况，(b) 真实 $f\left(x\right)$ 对应的 DCT 系数

3.1. 反演识别结果

在本章数值实验中，选择较为低频的二维 DCT 基函数，参数化基函数数量设为 120，即未知参数ξ的维数为$p=120$。EMVS 中待求参数初始值设置为$\xi=\mathbf{I}_P$，$\sigma=1$，$\theta=0.5$。“spike and slab”混合分层先验分布参数设定分别为$v_0=10000$，$v_1=2$，调整步长$\delta=0.001$。我们按照算法 1，在不同的预处理矩阵条件下，分别抽取了 30,000 个样本，其结果如图 2 所示。

从图中可以看出，经过预处理的 Langevin MCMC 方法可以得到通道的大致位置及形状，即使是在观测数据有限且聚集在边缘的情况下，而未经过预处理的 Langevin MCMC 方法的表现不尽如人意。由图 2 第二列的可以看出，在带预处理的 Langevin MCMC 方法下，其散点大致分布在$y=x$这条线上，绝大多数系数被截断为 0，即散点聚集在$y=0$的线上。重要的 DCT 基函数大多在模型中正确保留下来，即

相应系数没有被估计为 0。未预处理的 Langevin MCMC 算法，其中一些显著不为 0 的参数散点同样大致分布在 $y=x$ 这条线上，但其模型保留了过多变量，即过多的基函数没有被截断，这些基函数中包含了大量高频的 DCT 基向量。

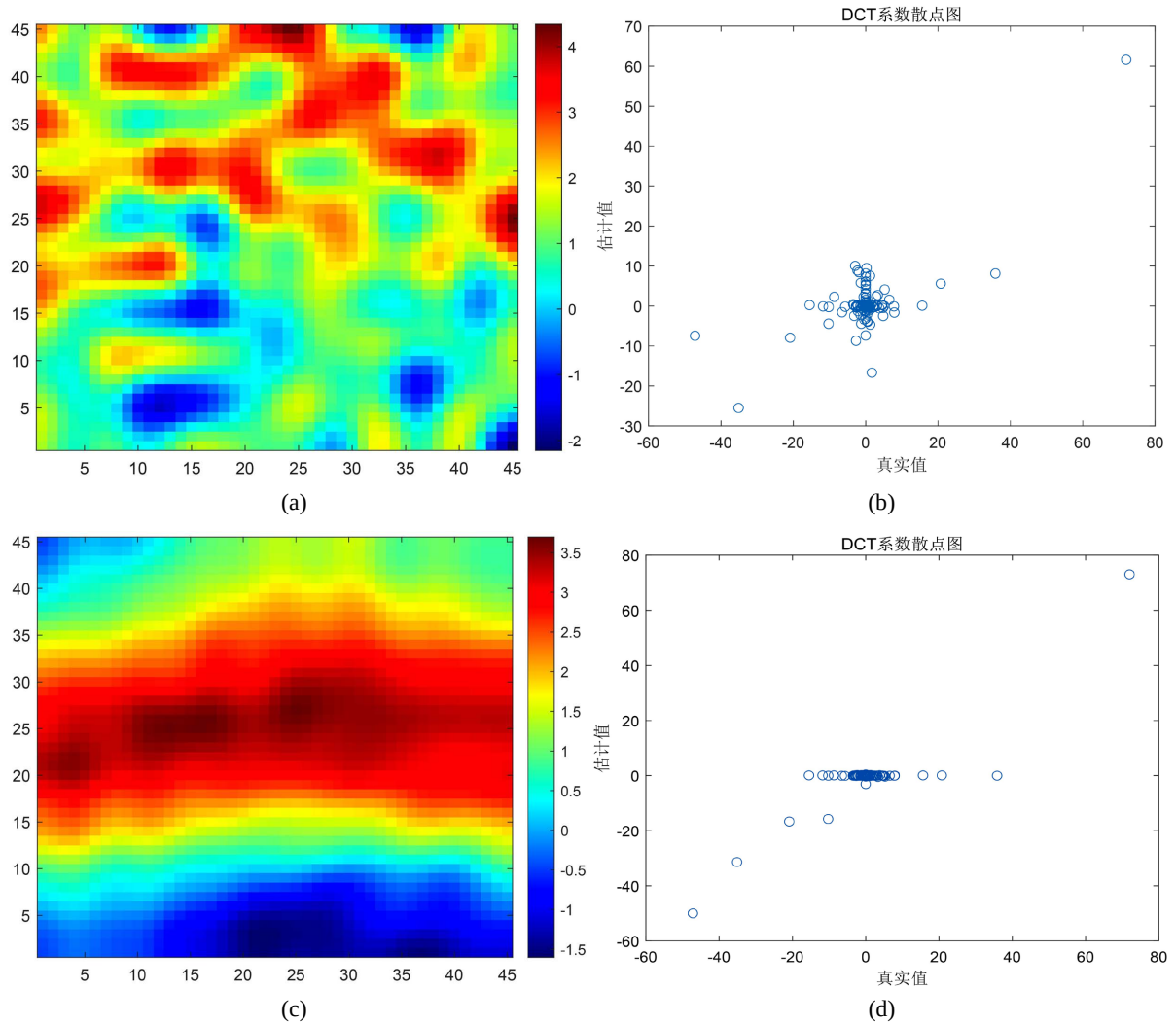


Figure 2. The first row is the results from the unpreconditioned Langevin MCMC method, the second row is the results from the inverse Hessian preconditioned Langevin MCMC method. The first column is the mean of $f(x)$, while the second column is the comparison between the estimated ξ and the reference ξ

图 2. 第一行为未经预处理的 Langevin MCMC 获得的结果，第二行为以逆 Hessian 矩阵为预处理矩阵的 Langevin MCMC 方法获得的结果。第一列为 $f(x)$ 的均值，第二列为 ξ 的估计值与参考值之间的对比图

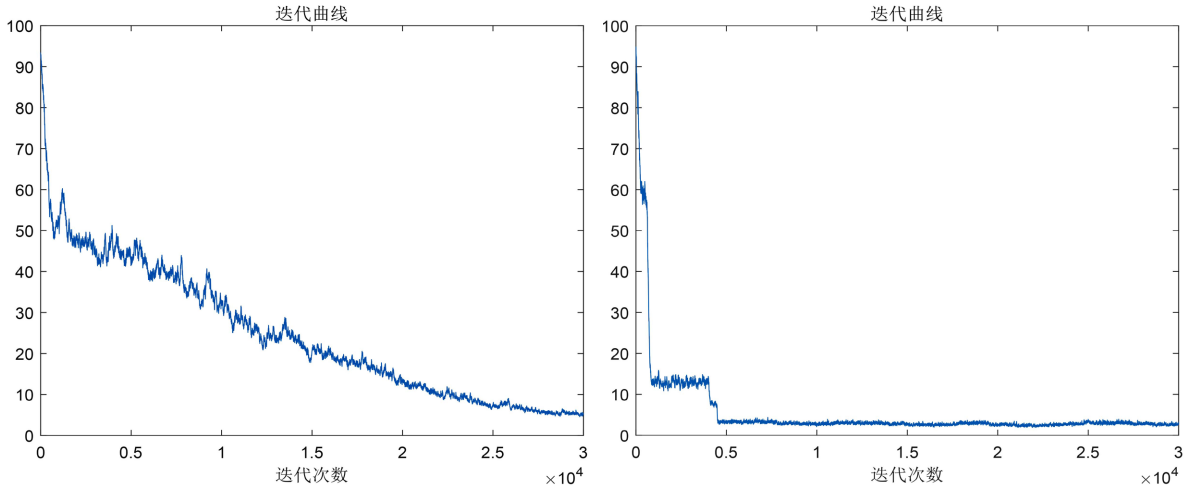
3.2. 收敛效果分析

按照算法 1 的原理和步骤，我们集合每次迭代中代数式

$$\arg\min\left\{\left\|\mathbf{d}-\mathbf{G}\left(\Phi\xi_k\right)\right\|^2+\left\|\mathbf{D}^{1/2}\xi_k\right\|^2\right\}$$

的值来刻画迭代曲线。迭代运算 30,000 次得到未经过预处理与经过预处理的方法各自相应的反演迭

代曲线如图 3 所示。我们发现无预处理算法的迭代曲线(图 3(a))始终没有进入稳定收敛区域，只在迭代初期曲线有明显剧烈下降。经过预处理的 Langevin MCMC 算法在前 5000 次迭代中已经急剧下降，并接近迭代最终值，这说明在算法中加入预处理的步骤可以有效提高计算的效率。

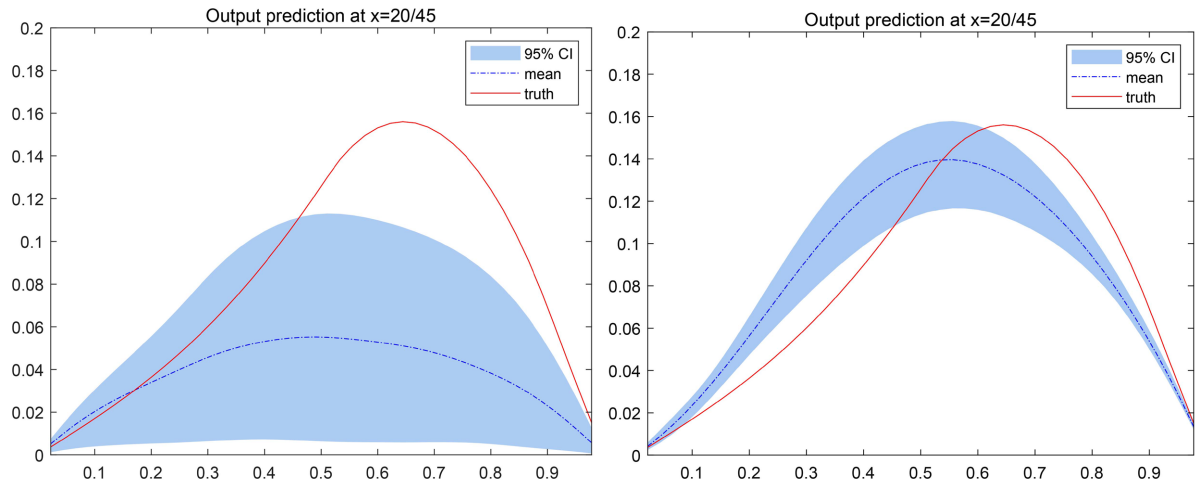


(a) (b)

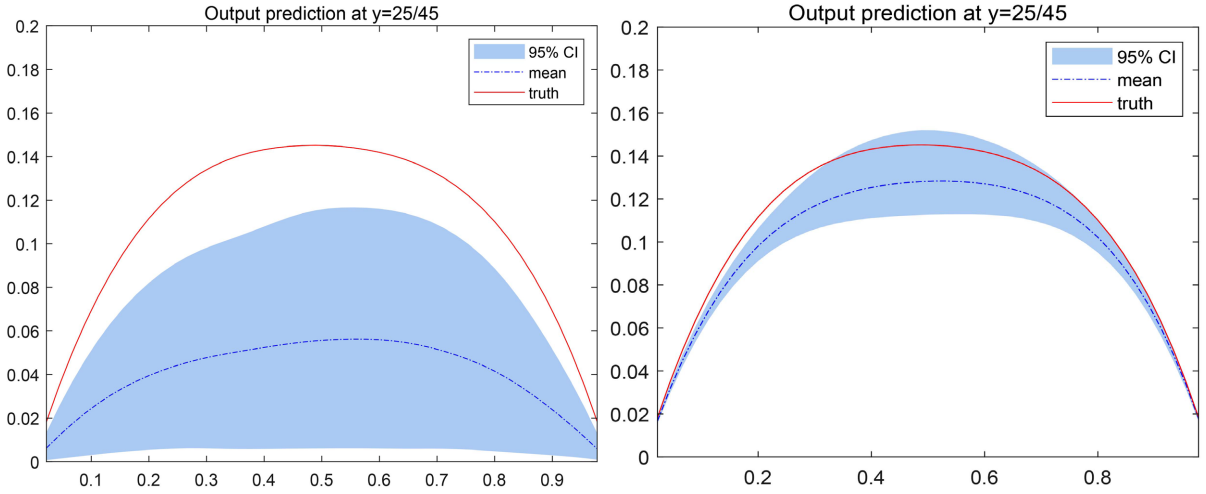
Figure 3. The plot of the objective function Q_1 against the iteration number, resulted from: (a) the unpreconditioned Langevin MCMC, (b) the Langevin MCMC with the inverse Hessian as the preconditioned matrix
图 3. 目标函数 Q_1 随迭代次数变化曲线图，结果来自于：(a) 未经预处理的 Langevin MCMC 方法，(b) 预处理矩阵为逆 Hessian 矩阵的 Langevin MCMC 方法

3.3. 模型预测

上一节我们发现经过预处理的 Langevin MCMC 方法对源函数的反演效果显著优于未经过预处理的 Langevin MCMC 方法，本节我们针对两种方法分别反演得到的待求参数后验样本，将参数 $\boldsymbol{\xi}$ 的样本代入到正演模型中，来得到相应的观测 $\mathbf{d}$ 的后验样本并对比。为了显示我们仅依靠网格边缘的部分观测数据，对参数的估计可靠性，尤其是在距离观测位置很远的中部，我们在图的中间位置任意选取 2 条函数线 $x = 20/45$ 和 $y = 25/45$，分别得到相应的预测输出值，同时与真实值作对比。



(a) (b)



(c) (d)

Figure 4. Predictions of the outputs for the unpreconditioned and inverse Hessian preconditioned Langevin MCMC method at (a, b) $x = 20/45$ (c, d) $y = 25/45$. The first column is from the unpreconditioned method, while the second column is from the inverse Hessian preconditioned method

图 4. 未经预处理和逆 Hessian 预处理 Langevin MCMC 方法下的输出变量分别在(a, b) $x = 20/45$ (c, d) $y = 25/45$ 处的预测情况。第一列为未经预处理方法的结果，第二列为经过预处理的结果

预测结果如图 4 所示，其中红色实线代表所选取位置上的真实值，蓝色虚线是将所提出方法反演得到的未知参数 ξ 的后验样本代入观测方程后得到的相应位置的输出样本均值，红色实线与蓝色虚线越靠近，说明数据预测结果越好。蓝色阴影区域是预测均值的 95%置信区间。可以看到，在任何预测位置上，经过预处理 Langevin MCMC 算法反演的未知参数 ξ 得到的预测结果其预测值与真实值都较为吻合，大多数都落在 95%置信区间内，因此认为我们的方法反演出来的参数估计结果可以用于模型预测。有预处理的 Langevin MCMC 方法所求未知参数预测得到的压力值较无预处理的 Langevin MCMC 方法所求未知参数压力预测值与真值更加接近。利用所提出的方法得到的对源项的反演结果可以用于模型预测，得到的预测值与真实值较吻合，预测精度较高。从预测结果来看，运用有预处理的 Langevin MCMC 方法所求未知参数求得的预测压力值与真值更加接近，结果精度更高。

4. 总结

本文在贝叶斯框架下，构建了一套地下水流非高斯源项场反演识别研究方法体系，该体系结合了数理方程正反演、离散余弦变化、再参数化、贝叶斯推断、Langevin MCMC、EMVS、预处理技术等多种理论与方法，综合运用理论分析和数值实验相结合的研究方式，针对地下水流非高斯参数场反演识别研究前沿中尚待解决的科学问题开展了系统性研究，丰富和拓展了地下水流非高斯源项场反演识别的理论基础，为实际地下油藏勘探、油气资源开发提供了重要参考。本文研究的假想例子为二维地下水非高斯源项场，条件较为简单。今后研究中，希望在更复杂的条件下进行识别研究：如地下水非线性渗透场，更大的观测误差，更少的观测数据等。

致 谢

欧娜感谢国家自然科学基金委 11901060，湖南省自然科学基金 2021JJ40557 以及湖南省教育厅优秀青年项目 22B0333 的支持；宋晓燕感谢国家自然科学基金委 12301551，湖南省自然科学基金 2022JJ40125 以及湖南省教育厅优秀青年项目 22B0635 的支持。

参考文献

[1] Kaipio, J. and Somersalo, E. (2006) Statistical and Computational Inverse Problems. Springer Science & Business Media, Berlin.

[2] Tarantola, A. (2005) Inverse Problem Theory and Methods for Model Parameter Estimation. Society for Industrial & Applied Mathematics, Berlin.

[3] Liu, J.S. (2008) Monte Carlo Strategies in Scientific Computing. Springer Science & Business Media, Berlin.

[4] Gamerman, D. and Lopes, H.F. (2006) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. CRC Press, New York. https://doi.org/10.1201/9781482296426

[5] Curtis, A. and Lomax, A. (2001) Tutorial: Prior Information, Sampling Distributions, and the Curse of Dimensionality. *Acta Chirurgica Italica*, **66**, 372-698.

[6] Cotter, S.L., Roberts, G.O., Stuart, A.M. and White, D. (2013) MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *Statistical Science*, **28**, 424-446. https://doi.org/10.1214/13-STS421

[7] Roberts, G.O. and Tweedie, R.L. (1996) Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, **2**, 341-363. https://doi.org/10.2307/3318418

[8] Gavalas, G.R., Shah, P.C., Seinfeld, J.H., *et al*. (1976) Reservoir History Matching by Bayesian Estimation. *Society of Petroleum Engineers Journal*, **16**, 337-350. https://doi.org/10.2118/5740-PA

[9] Karhunen, K. (1947) Uber linearemethoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae*, **37**, 3-79.

[10] Loeve, M. (1978) Probability Theory II. Springer-Verlag New York Inc., New York.

[11] Sarma, P., Louis, J.D. and Aziz, K. (2008) Kernel Principal Component Analysis for Efficient, Differentiable Parameterization of Multipoint Geostatistics. *Mathematical Geosciences*, **40**, 3-32. https://doi.org/10.1007/s11004-007-9131-7

[12] Jafarpour, B. and McLaughlin, D.B. (2009) Reservoir Characterization with the Discrete Cosine Transform. *SPE Journal*, **14**, 182-201. https://doi.org/10.2118/106453-PA

[13] Sahni, I. and Roland, R.N. (2005) Multiresolution Wavelet Analysis for Improved Reservoir Description. *SPE Reservoir Evaluation and Engineering*, **8**, 53-69. https://doi.org/10.2118/87820-PA

[14] Rockova, V. and George, E.I. (2014) EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, **109**, 828-846. https://doi.org/10.1080/01621459.2013.869223

[15] Jafarpour, B., Goyal, V.K., McLaughlin, D.B. and Freeman, W.T. (2010) Compressed History Matching: Exploiting Transform-Domain Sparsity for Regularization of Nonlinear Dynamic Data Integration Problems. *Mathematical Geosciences*, **42**, 1-27. https://doi.org/10.1007/s11004-009-9247-z

[16] Wang, Y.T., Deng, W. and Lin, G. (2021) Bayesian Sparse Learning with Preconditioned Stochastic Gradient MCMC and Its Applications. *Journal of Computational Physics*, **432**, 110134. https://doi.org/10.1016/j.jcp.2021.110134

[17] Deng, W., Zhang, X., Liang, F.M. and Lin, G. (2019) An Adaptive Empirical Bayesian Method for Sparse Deep Learning. *Advances in Neural Information Processing Systems* 32.

[18] Andrew, M.S., Voss, J. and Wilberg, P. (2004) Conditional Path Sampling of SDEs and the Langevin MCMC Method. *Communications in Mathematical Sciences*, **2**, 685-697. https://doi.org/10.4310/CMS.2004.v2.n4.a7

[19] Martin, J., Lucas, C.W., Burstedde, C., *et al*. (2012) A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion. *SIAM Journal on Scientific Computing*, **34**, 1460-1487. https://doi.org/10.1137/110845598