水 利 学 报
2009 年 8 月 SHUILI XUEBAO 第 40 卷 第 8 期

文章编号:0559-9350(2009)08-1019-05

# 改进的 MCMC 方法及其应用

朱嵩[1],毛根海[1],刘国华[1], 黄跃飞[2]

(1. 浙江大学 建筑工程学院,浙江 杭州 310058; 2. 清华大学 水利水电工程系,北京 100084)

**摘要**:概率反演中,马尔科夫链蒙特卡罗是一类重要的后验概率抽样方法,但由于该算法的搜索往往会陷入局部最优解,因而限制了其在具有非唯一解反问题中的应用。鉴于此,本文对基于 Metropolis-Hastings 算法的多链搜索的方法进行了改进,改进后的方法可以根据搜索结果实时调整链的个数,因而可以在搜索到尽可能多的解的同时节省了多链搜索的时间。最后将该算法应用于一个地下水污染源反问题的求解,计算结果表明改进后的算法对求解非唯一性反问题具有较好的效果。

**关键词**:马尔科夫链蒙特卡罗;概率反演;Metropolis-Hastings 算法;非唯一性;环境水力学

**中图分类号**:O242; TV13 **文献标识码**:A

## 1 研究背景

一般而言,环境水力学正问题主要研究地表及地下水中污染物的输移、扩散和转化规律,建立相关的分析计算方法,确定污染物浓度的时空分布及其应用[1]。与此对应,环境水力学反问题是指根据有限且离散的实测水动力、水质数据来估计环境水力学模型参数、边界条件、初始条件以及污染源位置和强度等信息。与正问题的适定性相反,环境水力学反问题是一种不适定问题,主要表现为解的不唯一性,这给反问题的求解带来了较大的困难[2]。

目前,反问题求解方法主要包括正则化方法、最优化方法、概率统计方法等[3]。在环境水力学反问题研究领域中,这些方法都得到了应用[4-9]。然而针对环境水力学中广为存在的测量噪声,以及先验信息等不确定性因素,概率统计类方法相比较优化类方法能更好地描述和求解此类问题[10]。贝叶斯推理(Bayesian inference)是处理非线性、不确定性系统反问题一种行之有效的概率反演方法,它通过对未知参数的后验概率进行抽样来获得参数的估计。马尔科夫链蒙特卡罗(Markov chain Monte Carlo,MCMC)目前是贝叶斯推理的标准抽样方法,但由于非线性不确定性系统的复杂性及 MCMC 算法中计算参数设置的人为性(如计算步长和计算终止条件的选择等),抽样结果是否具有对后验概率分布的代表性与 Markov 链设计者的经验有很大关系。由于对一次计算结果的可靠性很难做出判断,因而采用多链搜索方法对其进行改进是一种有效的办法,例如 Metropolis-coupled MCMC 算法[11]等。

本文亦对基于 Metropolis-Hastings 算法的 MCMC 抽样方法进行改进,提出了一种动态多链方法(Dynamic Multi-chain Metropolis-Hastings,DMMH),该方法在不改变 Metropolis-Hastings 算法核心机制上能通过增加搜索结果对链数目的反馈机制,从而保证每条 Markov 链收敛性的同时提高抽样的全局能力和多链抽样的效率。最后将该算法应用于求解地下水污染源估计反问题,计算结果表明改进后的算法显著增强了 MCMC 搜索的全局能力,同时降低了多链搜索的总时间。

收稿日期:2008-06-16

基金项目:973 课题(2005CB724202);国家自然科学基金项目(50609024);浙江省自然科学基金(Y506138)

作者简介:朱嵩(1981—),男,安徽人,博士后,主要从事环境水力学反问题研究。E-mail:migao@zju.edu.cn

## 2 MCMC 算法及其改进

MCMC 是利用 Markov 链机制探索状态空间以生成样本的方法，这种机制能够保证 Markov 链能花更多的时间在最重要的区域，尤其它能够被构造，以致它产生的样本能够模仿目标分布的样本[12]。Markov 链的重要特性是无后效性，它指事物本阶段的状态只与前一个阶段的状态有关，而与以前其他任何阶段的状态无关。Metropolis-Hastings 算法[13—14]是应用最为广泛的 MCMC 抽样方法。大多数实际的 MCMC 算法可以解释成 Metropolis-Hastings 算法的特例或扩展。关于 Metropolis-Hastings 算法可参见文献[12—14]，限于篇幅这里不再赘述。

由于计算时间有限和随机游走设计的不当，对于一个较为复杂的参数空间，一个基于 Metropolis-Hasting 算法的实际随机游走往往不能对其充分搜索。因此对于一个具有潜在非唯一解的反问题，采用一个陷入局部最优解的随机游走产生的样本来对反问题的解进行评价是不全面的，甚至可能是误导的。如图 1 所示，对于一个多峰值的概率密度函数，传统的 Metropolis-Hastings 算法往往会陷入一个局部高密度区域。

为了提高 MCMC 抽样的全局能力，本文提出动态多链的抽样方法 DMMH。算法如图 2 所示：

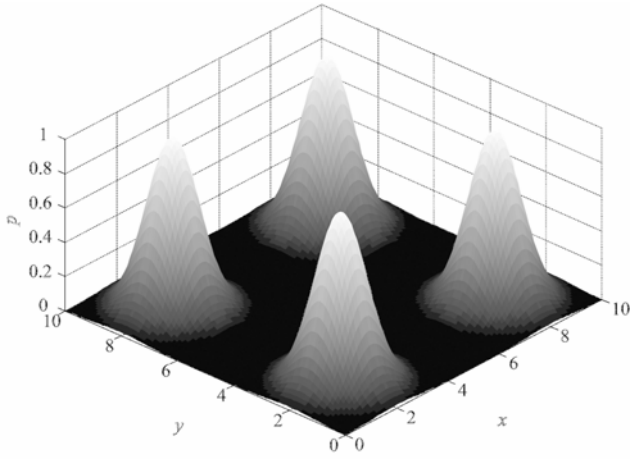


图 1 一个多峰值的后验概率密度函数

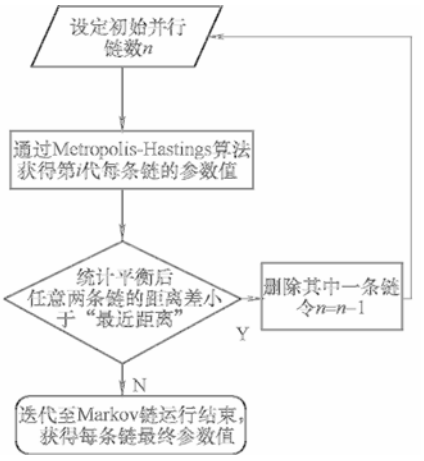


图 2 DMMH 算法流程

该算法对经典的 Metropolis-Hastings 算法的改进如下：(1)设置了多条初始点不同的 Markov 链对整个后验空间进行搜索，起始点的位置在参数的先验空间内随机产生，这样尽可能提高了早期 Markov 链的多样性，进而提高了抽样的全局能力；(2)在每条链进入统计平衡阶段后，通过判断链与链之间的接近程度(根据反演参数的统计特征)来实时调整(减少)多链的数目，以尽可能提高计算效率。

初始链数的数目亦与局部最优解的数目相适应，太大则会影响早期的计算效率，太小则多链搜索的优势难以体现。链与链之间的接近程度，可以由当前代每条链的当前值或统计平均值的欧氏距离等来表达。例如对于参数 $\boldsymbol{X}$ 第 $i$ 代时第 $j$ 条链和第 $k$ 条链，其距离可以表达为 $\| \boldsymbol{X}_i^{(j)} - \boldsymbol{X}_i^{(k)} \|_2$。

## 3 算例验证

为验证 DMMH 算法的可靠性和相对于经典 Metropolis-Hastings 算法的优点，本文选取了一个地下水污染源识别问题作为算例。为了便于说明问题，假设水动力—水质场的参数均已知，而污染源的位置未知。

设在无限展布、水平、等厚的均质含水层中存在着均匀单向流动，取 $x$ 轴方向为流动方向，在 $t=0$ 开始在污染源处($x_0$，$y_0$)向含水层连续注入含示踪剂流体。

根据文献[15]并加以推导，可得在任意一点($x$，$y$)上，$t$ 时刻的污染物浓度值，如式(1)：

$$c(x,y,t)=\frac{c_0 q}{4\pi\sqrt{D_L D_T}}\exp\left\{\frac{V(x-x_0)}{2D_L}\right\}\int_{\frac{a}{4t}}^{\infty}\exp\left\{-u-\frac{ab}{u}\right\}\frac{\mathrm{d}u}{u} \tag{1}$$

式中：$c_0$ 为示踪剂浓度；$q$ 为示踪剂的注入速率；$D_L$、$D_T$ 分别污染物纵向弥散系数和横向弥散系数；$V$ 为水体在 $x$ 轴方向上的流速；$a$、$b$ 分别为两个系数：$a=\frac{(x-x_0)^2}{D_L}+\frac{(y-y_0)^2}{D_T}$，$b=\frac{V^2}{4D_L}$。算例中流动及扩散参数如下：流速 $V=0.1$m/s，示踪剂注入速率点源强度 $c_0=100$mg/L，纵向扩散系数 $D_L=1$m²/s，横向扩散系数 $D_T=0.3$m²/s，$q=1$m²/s。

设该问题唯一的浓度观测点为 $P$(300m,150m)，由于对称性可知，污染源的潜在位置关于过 $P$ 点的水平轴对称，因而该问题是一个具有非唯一解的反问题。设污染源位置($x_0$，$y_0$)的两个“真值”为(200m,100m)及(200m,200m)。通过式(1)可以计算得到测点 $P$ 上的浓度值作为观测值，共取了 $t=1$、2、3、4、5 和 6min 时的浓度计算值，如表 1 所示。此外，假设测量噪声为白噪声 $N(0,\sigma^2)$，$\sigma=1\mathrm{E}-6$。

表 1　观测点上污染物浓度“测量值”

| $t$/min | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $c$/(mg/L) | 0.000 133 173 | 0.010 994 5 | 0.054 178 6 | 0.126 884 | 0.217 631 | 0.317 398 |

下面分别采用 Metropolis-Hastings 算法和 DMMH 算法对污染源进行反演计算。

**3.1　应用 Metropolis-Hastings 算法**　模型参数的先验分布为 $x\in[150,250]$m，$y\in[50,250]$m 上的均匀概率分布。随机游走的 proposal 分布为 $g(x^*\mid x^{(i)})=U(x^{(i)}-\mathrm{step},x^{(i)}+\mathrm{step})$，步长 step 为参数先验范围的 40%，即参数 $x$ 的搜索步长为 40m，$y$ 的搜索步长为 80m。迭代进行 1 000 次。图 3 为两个参数的迭代曲线，从图中可以看出，Metropolis-Hastings 算法只搜索到一个污染源位置(200m,100m)附近。

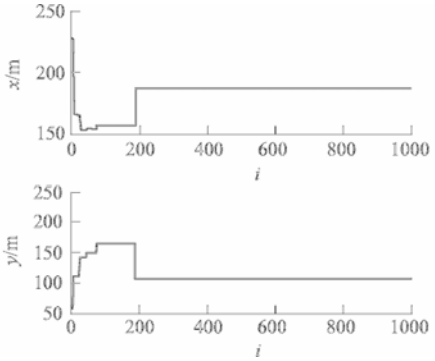


图 3　污染源位置迭代曲线
(Metropolis-Hastings 算法)

**3.2　应用 DMMH 算法**　其他条件不变，采用 DMMH 算法进行计算。起始链数取为 4 条，采用当前值作为 2 条链接近的度量方式，当式(2)所示条件满足时删除编号较小的那条链：

$$\begin{cases}\mid x_i^{(j)}-x_i^{(k)}\mid<50\mathrm{m}\\ \mid y_i^{(j)}-y_i^{(k)}\mid<50\mathrm{m}\\ i\geqslant 200\end{cases} \tag{2}$$

图 4—7 为 4 条链的迭代曲线，表 2 为各链的运行过程及迭代 2 000 步时的搜索结果。从图 4—7 和表 2 中可以看出，DMMH 算法搜索到了 2 个污染源的潜在位置(200m,100m)及(200m,200m)附近。若采用迭代运行终点计算值作为污染源的估计值，则两条链的搜索的相对误差分别 19.5%和 10.3%(见表 2)。需要指出，本算例中由于测量噪声相当小，使得参数“真值”影响区域只占整个后验参数空间中一个极窄的区域，因而随机游走进行得相对较慢。从图 6—7 中我们也可以看出，在迭代过程中搜索曲线在相当长的范围内没有变化。尽管如此，MCMC 搜索机制能够保证搜索向高概率密度区域(真值附近)方向搜索。若要进一步提高搜索精度，需要进一步增加 Markov 的迭代步数。

表 2　Markov 链运行状态及搜索结果

| 链编号 | 运行状态 | 搜索到污染源的位置 | 搜索值距真值的距离 | 相对先验范围的误差 |
|---|---|---|---|---|
| 1 | 与链 2 较接近，200 步时被删除 | | | |
| 2 | 与链 4 较接近，201 步时被删除 | | | |
| 3 | 迭代至终点 | (182.5m,108.6m) | 19.5m | 19.5% |
| 4 | 迭代至终点 | (219.6m,206.3m) | 20.6m | 10.3% |

跟踪 4 条链的运行状态，可以发现当迭代至 200 步时，链 1 和链 2 接近，因而链 1 被删除；迭代至 201 步时，链 2 和链 4 接近，因而链 2 被删除。假设每条链的总计算时间为 T，则 DMMH 算法的计算时间

为 200*T*?2000+201*T*?2000+*T*+*T*=2.2005*T*,约为 4 条链数固定条件下多链搜索时间(4*T*)的 55%,表明 DMMH 算法相对于常规的多链搜索具有较高的计算效率。

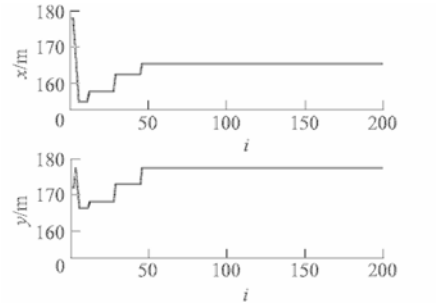


图 4　链 1 中参数的迭代曲线(DMMH 算法)

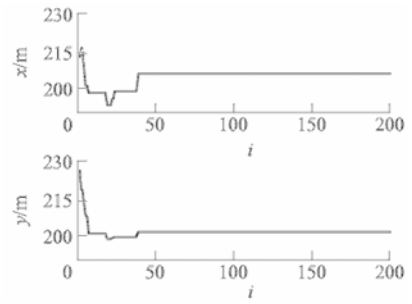


图 5　链 2 中参数的迭代曲线(DMMH 算法)

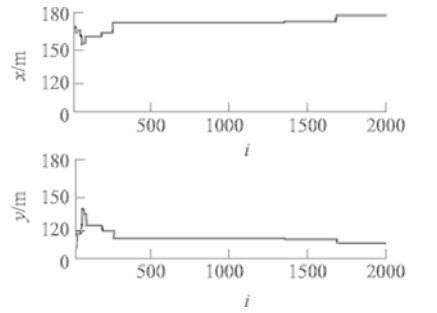


图 6　链 3 中参数的迭代曲线(DMMH 算法)

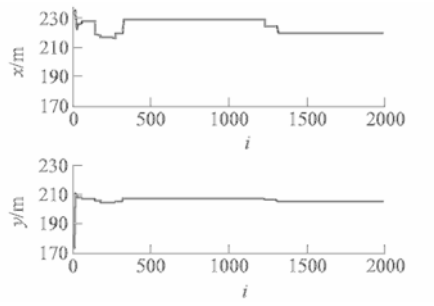


图 7　链 4 中参数的迭代曲线(DMMH 算法)

## 4　结论与讨论

本文针对具有非唯一解的反问题,在经典 Metropolis-Hastings 算法的基础上提出了一种动态多链的蒙特卡罗方法 DMMH。该方法由于设置了多链的搜索结果与链的数目之间的反馈机制,因而实现了 Markov 链的搜索的时间复杂度与反问题复杂性(主要表现在非唯一解的特性)之间的一个较为合适的平衡。对于具有潜在非唯一解的反问题,DMMH 算法将在一个合适的计算时间内搜索到尽可能多的非唯一解。

由于参数的后验空间往往是相当复杂的,因而在有限的时间内寻找到所有的可行解是相当困难的,充分利用高性能计算是解决此类问题的一个十分重要的途径。

**参　考　文　献:**


[1] 李玉梁,李玲.环境水力学的研究进展与发展趋势[J].水资源保护,2002(1):1—6.

[2] Kirsch Andreas. An Introduction to the Mathematical Theory of Inverse Problems[M]. New York: Springer, 1996.

[3] 王彦飞.反演问题的计算方法及其应用[M].北京:高等教育出版社,2007.

[4] Isakov Victor, Kindermann Stefan. Identification of the diffusion coefficient in one-dimensional parabolic equation[J]. Inverse Problems, 2000, 16(3): 665—680.

[5] 闵涛,周孝德,冯民权.非线性布西尼斯克方程的直线解法及渗透系数反演计算[J].水利学报,2004(7):21—25.

[6] Liu Shuming, Butler David, Brazier Richard, et al. Using genetic algorithms to calibrate a water quality model[J]. Science of the Total Environment, 2007, 374, (2—3), 15: 260—272.

[7] 闵涛,周孝德,冯民权.河流水质多参数识别反问题的演化算法[J].水利学报,2003 (10):119—123.

[8] Liu Yong, Yang Pingjian, Hu Cheng, et al. Water quality model for load reduction under uncertainty: A Bayesian

approach[J]. Water Research, 2008, 42 (13): 3305—3314.
[9] 朱嵩.基于贝叶斯推理的环境水力学反问题研究[D].杭州:浙江大学,2008.
[10] Albert Tarantola. Inverse problem theory and methods for model parameter estimation[M]. Philadelphia: SIAM, 2004.
[11] Geyer C J. Markov chain Monte Carlo maximum likelihood[C]? In Keramidas (ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium Interface. Washington: Institute of Mathematical Statistics, 1991:156—163.
[12] Andrieu Christophe, De Freitas Nando, Doucet Arnaud, et al. An introduction to MCMC for Machine Learning[J]. Machine Learning, 2003, 50: 5—43.
[13] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. Equations of state calculations by fast computing machines[J]. Journal of Chemical Physics, 1953, 21: 1087—1091.
[14] Hastings W K. Monte Carlo sampling methods using Markov chains and their Applications[J]. Biometrika, 1970, 57: 97—109.
[15] 孙讷正.地下水污染一数学模型和数值方法[M].北京:地质出版社,1989:60—61.

## Improved MCMC method and its application

ZHU Song[1], MAO Gen-hai[1], LIU Guo-hua[1], HUANG Yue-fei[2]
(1. *Zhejiang University, Hangzhou* 310058, *China*; 2. *Tsinghua University, Beijing* 100084, *China*)

**Abstract**: A multi-chain sampling method based on Metropolis-Hastings algorithm was used to improve the Markov Chain Monte Carlo (MCMC) method in order to prevent from trapped into the local optimal solutions that often occur to probability inversion by using current MCMC algorithm. The improved MCMC method can adjust the number of chains according to the feedback results from sampling process in real time, so that it can search out the non-unique solutions as much as possible while saving the time of multi-chain search. As an example an inverse problem of groundwater flow was solved by using the improved MCMC algorithm. The computational results indicate the improved method performs well in solving inverse problems with non-unique solutions.

**Key words**: Markov Chain Monte Carlo; probability inversion; Metropolis-Hastings algorithm; non-uniqueness; environmental hydraulics

(责任编辑:韩　昆)