基于 MCMC 法的水质模型参数不确定性研究

王建平, 程声通, 贾海峰*
(清华大学环境科学与工程系, 北京 100084)

摘要: 参数识别是数学模型应用的前提. 鉴于常用贝叶斯离散化方法在搜索复杂模型参数后验分布时的计算限制的原因, 本文引入了 MCMC 采样法. 为考察 MCMC 法对参数后验分布的搜索性能和效率, 进行了 2 个案例研究. 结果表明, MCMC 法对参数后验分布的搜索, 无论是搜索性能还是搜索效率, 均表现出了独特的优越性. 同时, Gelman 收敛判别准则计算表明, MCMC 采样序列均能稳定收敛到参数的后验分布上. 可见, MCMC 法适用于复杂环境模型的参数识别和不确定分析研究.

关键词: 马尔科夫链蒙特卡罗法; 水质模型; 不确定分析; 参数识别
中图分类号: X192 **文献标识码**: A **文章编号**: 0250-3301(2006)01-0024-07

Markov Chain Monte Carlo Scheme for Parameter Uncertainty Analysis in Water Quality Model

WANG Jian-ping, CHENG Sheng-tong, JIA Hai-feng
(Department of Environmental Science and Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Parameter identification plays an important role in environmental model application. Markov Chain Monte Carlo method was introduced to estimate parameter uncertainty, since usual Bayes discrete methods were not applicable to produce posterior distribution of complicated environmental model due to the limit of computation. In order to study the performance and efficiency of MCMC, two case studies were used. Results indicate that, either sampling performance or sampling efficiency, MCMC method both has its special advantages in producing posterior distribution. Moreover, results of Gelman convergence diagnostics indicate that sampling sequence can converge to a stationary distribution. A key finding was that the MCMC scheme presented herein provided a powerful means of parameter identification and uncertainty analysis.

Key words: MCMC; water quality model; uncertainty analysis; parameter identification

一个能反映客观实际的数学模型, 只有在它得到合理的参数数值之后, 才具有生命力[1]. 随着水质模型在环境规划与管理中不断深入和广泛地应用, 模型结构复杂性急剧增长, 模型参数在高维空间表现出了复杂的相关性结构和低灵敏度特征, 从而直接导致了优化后验参数的识别问题[2,3]. 传统参数识别中仅局限于参数优化算法效率和精度等方面的研究已经不能满足理论与实践的需要, 相对于观测数据和模型参数而言, 基于现有科学认知体系构建的模型结构是建模过程中不确定性的根本来源[4]. 然而由于缺乏深入研究结构不确定性的理论基础和有效技术手段, 模型结构往往只能通过识别参数后验分布统计规律间接地得到验证, 因而参数不确定性分析研究就显得格外重要.

1 参数不确定性分析

伴随着模型的发展和研究人员对模型结构特征的认识, 参数识别研究基本围绕 2 种思路展开[5]: ①基于优化思想的参数识别思路, 即传统的参数识别思路; ②基于不确定分析的参数识别思路. 可识别性是前 1 种参数识别思路的前提. 然而, 水质模型参数识别属于复杂非线性问题, 参数响应曲面存在很多凹谷和平坦区域, 有大量局部极小点, 优化算法往往只能得到局部最优解. 参数越多, 参数响应曲面的非线性度越高[6], 参数不可识别性问题越严重. 科学家们采用各种方法克服模型的不可识别性问题, 最常见的方法是将模型的部分参数(通常是那些灵敏度较低的参数)预先采用先验知识固定, 从而减少待识别参数个数; 或者对模型进行简化, 去除那些不能被观测数据识别的子过程[7]. 这种参数识别思路的弊端是显而易见的. 简化模型意味着模型机理的减少, 这往往是研究人员不希望看到的, Reichert 指出, 一个唯一确定的模型用于预测, 模型结构的简化可能导致模型预测不确定性的低估, 尤其是当一个过程在识别阶段不重要, 而在预测阶段变得重要时更是如此[7].

为了克服和解决参数不可识别的问题, 基于贝

收稿日期: 2004-10-20; **修订日期**: 2004-12-14
基金项目: 国家自然科学基金资助项目(50209007)
作者简介: 王建平(1977~), 男, 博士研究生, 主要从事环境系统分析的理论与应用研究, E-mail: wangjp@tsinghua.org.cn
* 通讯联系人, jhf@tsinghua.edu.cn

叶斯理论的不确定性参数识别思路应运而生. Tiwari 最早将贝叶斯理论用于生态模型的参数识别[8], 随后 Hornberger 与 Spear 提出了 RSA 方法[9,10], Beven 提出了 GLUE 方法[6]. 与传统的参数识别思路相比, 参数不可识别性对贝叶斯统计法来说不再是一个问题, 因此对于具有不可识别参数的模型, 贝叶斯方法无疑是很好地选择[5,11]. 在统计推断中使用先验分布的方法就是贝叶斯方法, 也就是说, 是否使用先验分布是区分贝叶斯统计和非贝叶斯统计的标志. 非贝叶斯理论在做统计推断时只依据 2 类信息, 即模型结构信息和数据信息. 而贝叶斯统计除了依据以上 2 类信息, 还要利用另一类信息, 即未知参数的分布信息. 由于这类信息是在获得实际观测数据以前就有的, 因此一般称为先验信息(A Prior Information). 贝叶斯统计要求这类信息能以未知参数的统计分布来表示, 这个概率分布就称为先验分布. 根据贝叶斯理论, 参数的先验分布、样本信息和后验分布具有如下的关系[12]:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \quad (1)$$

$p(\theta \mid y)$ 是参数的后验分布密度, $p(\theta)$ 是参数的先验分布密度, $p(y \mid \theta)$ 体现了在现有的数据条件下参数的似然度信息, $p(y)$ 为比例常数.

贝叶斯方法模式简单, 概率形式优美. 然而, 它的数值解法并非总是容易的、直接的. 实际应用中需进行随机变量的离散化, 如随机采样算法 RSA 法和 GLUE 法均属这一类算法. 基于随机采样的统计方法, 可以获得模型参数的后验分布, 而不再是一组单一的最优参数, 在一定程度上避免了由于"最优"参数失真而带来的决策风险. 但是, 由于参数的产生是随机的, 属于"盲"搜索, 当参数较多时, 要想获得具有一定代表性的采样点数, 就必须进行大量的采样, 计算量随参数的增多呈指数增长. 总之, 贝叶斯方法用于参数识别的思想已得到不同领域研究工作者的广泛认可, 但是由于其巨大的计算量, 在实际中很难推广应用, 这些离散贝叶斯方法只适合参数个数较少的情况, Jorgensen 指出, 当参数个数超过 3~5 个时, 参数识别过程将非常耗时[13].

离散贝叶斯方法应用的主要障碍出在计算上, 即使采用高性能计算机进行模拟, 也面临着计算复杂性的问题[5,7]. 20 世纪 90 年代, 研究人员将马尔科夫链蒙特卡罗法(Markov Chain Monte Carlo, MCMC)[14]引入到参数不确定性研究中, 用于待估参数的贝叶斯分布采样, 以估计参数的后验分布. 由于这种方法的应用, 使得随机模拟在很多领域的计算中, 显示出巨大的优越性, 相比 Monte Carlo 法, MCMC 法可大大降低计算量[14]. 本文将应用 MCMC 法进行模型参数的不确定性分析研究.

2　MCMC 法

自 1907 年苏联数学家 Markov 提出马尔科夫链的概念以来, 经过世界各国几代数学家的相继努力, 目前马尔科夫链已成为内容十分丰富、理论上相当完整的数学分支. 马尔科夫链理论已成为强有力的数学工具, 广泛应用于物理、化学、生物、天文、地质、气象、计算机、通信等众多领域[15]. 马尔科夫链有严格的数学定义, 其直观意义可理解为: 随机系统中下一个将要达到的状态, 仅依赖于目前所处的状态, 而与以往所经历过的状态无关[15].

MCMC 法用于模型参数不确性分析的研究是近年来才发展起来的一种方法. Smith 于 1993 年[16]提出利用 MCMC 法获取参数的后验分布, 随后 Tierney(1994)[17] 与 Chib(1995)[18] 相继发表了 MCMC 法用于参数不确定性分析的理论研究成果.

2.1　采样算法

常用的 MCMC 采样算法有: Metropolis-Hastings 算法[17]和自适应 Metropolis 算法[19].

2.1.1　Metropolis-Hastings 算法

Metropolis-Hastings 算法(简称 M-H 法)是基于贝叶斯推理框架下描述参数不确定性的最早、最通用的一类 MCMC 采样器. M-H 算法获取参数后验分布的各态历经(Ergodicity)和收敛特性在文献中已有详细研究, 并给出了算法收敛的几何条件[20]. 有关 M-H 法的描述详见参考文献[15,17]. 应用 M-H 采样器的关键是确定参数的推荐分布(Proposal distribution)和参数相关性的处理. 对于复杂模型来说, 参数先验信息较少, 确定参数的希望区域(高概率密度区域)非常困难, 参数推荐分布的选择会带来很大的初始不确定性, 导致收敛速率缓慢. 为改进 MCMC 采样器的搜索效率, 人们希望算法随采样过程自适应地调整参数的推荐分布, 这样, 自适应的 Metropolis 算法应运而生. 自适应的 Metropolis 算法可以有效地解决 M-H 法存在的问题.

2.1.2　自适应 Metropolis 算法

自适应 Metropolis 算法(Adaptive Metropolis, AM)是 Haario 在 2001 年提出的一种改进 MCMC 采样器[19]. 相比传统 M-H 算法, AM 算法不再需要事先确定参数的推荐分布, 而是由后验参数的协方

差矩阵来估算. 后验参数的协方差矩阵在每一次迭代后自适应地调整. 这样, 第 i 步参数的推荐分布为均值 θ_i, 协方差 C_i 的多元正态分布. 协方差计算公式如式(2), 在初始 i_0 次迭代中, 协方差矩阵 C_i 取固定值 C_0, 之后自适应更新.

$$C_i = \begin{cases} C_0 & i \leqslant i_0 \\ s_d \mathrm{Cov}(\theta_0, \cdots, \theta_{i-1}) + s_d \varepsilon I_d & i \geqslant i_0 \end{cases} \quad (2)$$

其中, ε 为一个较小的数, 以确保 C_i 不成为奇异矩阵; s_d 为一个比例因子, 依赖于参数的空间维度 d, 以确保接受率在一个合适范围内; I_d 为 d 维单位矩阵. 第 $i+1$ 次迭代, 由公式(2)可推出协方差计算公式(3), 可以看到, 计算量很小.

$$C_{i+1} = \frac{i-1}{i} C_i + \frac{s_d}{i}(i\overline{\theta}_{i-1}\overline{\theta}_{i-1}^T - (i+1)\overline{\theta}_i\overline{\theta}_i^T + \theta_i\theta_i^T + \varepsilon I_d) \quad (3)$$

其中, $\overline{\theta}_{i-1}$ 和 $\overline{\theta}_i$ 为前 $i-1$ 和 i 次迭代参数的均值.

AM 算法的搜索流程如下:

(1) 初始化, $i=0$;

(2) 状态随机产生和接受;

(a) 利用公式(2)计算 C_i;

(b) 产生推荐参数值 $\theta^* \sim N(\theta_i, C_i)$;

(c) 计算接受概率 α;

$$\alpha = \min\left[1, \frac{p(y \mid \theta^*)p(\theta^*)}{p(y \mid \theta_i)p(\theta_i)}\right] \quad (4)$$

参数物理意义同前.

(d) 产生随机数 $u \sim U[0,1]$;

(e) 若 $u < \alpha$ 接受 $\theta_{i+1} = \theta^*$, 否则 $\theta_{i+1} = \theta_i$;

(3) 重复(a) ~ (e)直到产生足够的样本为止.

AM 算法的最大优点是推荐分布随计算过程自动更新, 不再需要事先指定. 同时, 相比传统 M-H 算法, 参数同时更新, 不再需要分组更新, 计算量大大减少. 基于以上优点, 本研究将采用 AM 算法来搜索参数后验分布.

2.2 MCMC 法采样设计

2.2.1 算法参数选择

初始化阶段("Burn in" period): 受模型参数初值 θ_0 影响的初始迭代序列 θ_i, 即 θ_i 后的采样点将收敛到参数的后验分布上. 初始化阶段在统计分析中必须去除, 以消除初值的影响. 初始化阶段长度的确定是 MCMC 法应用的一个难点. 通常根据统计结果确定.

参数初始协方差矩阵 C_0 和初始迭代次数 i_0 的确定. AM 算法中 C_0 的指定没有严格要求, 通常根据参数的先验分布来确定.

ε 和 s_d 的确定. 本研究中 $\varepsilon = 10^{-5}$, $s_d = (2.4)^2/d$, d 为参数个数[19].

参数先验分布的确定. 本研究采用均匀分布.

似然度函数的确定. 似然度函数的确定有很多方法[6], 本研究似然度函数的形式为:

$$p(y \mid \theta) = \frac{1}{\prod_{k=1}^{L} E_k} \quad (5)$$

其中, $E_k = \dfrac{\sum_{i=1}^{n} \mid C_{ki} - \hat{C}_{ki} \mid}{\sum_{i=1}^{n} C_{ki}}$

式中, $k = 1, 2, \cdots, L$; L 为系统模拟的水质状态变量数; $i = 1, 2, \cdots, n$; n 为监测次数; E_k 为第 k 变量的相对误差; C_{ki} 为第 i 次监测第 k 变量的浓度模拟值; $\hat{C}_{ki}$ 为第 i 次监测第 k 变量的浓度观测值.

2.2.2 收敛判断准则

MCMC 研究的一个重要任务是判断采样序列是否收敛到参数后验分布. 理论上, 一个各向同性的采样器在 $t \to \infty$ 时一定收敛, 然而在实际应用中并非如此. Gelman[21] 在 1992 年提出了一种定量的收敛诊断指标 $\sqrt{R}$, 称为比例缩小得分(Scale Reduction Score), 计算方法如下:

$$\sqrt{R} = \sqrt{\frac{g-1}{g} + \frac{q+1}{q \cdot g}\frac{B}{W}} \quad (6)$$

其中, g 为每一参数采样序列的迭代次数; q 为用于评价的序列数; B/g 为 q 个序列的平均值的方差; W 为 q 个序列的方差的平均值. 计算每一参数的比例缩小得分 $\sqrt{R}$, 若接近于 1 表示参数收敛到了后验分布上.

Gelman 提出的方法为多序列对比方法, 研究中考察单序列是否稳定的方法有平均值法和方差法, 即考察迭代过程中的平均值和方差是否稳定. 显然, 单序列评价方法不能判别序列是否全局收敛.

3 案例研究

3.1 双峰概率密度函数(案例 1)

为考察 MCMC 法对参数后验分布的搜索效率和性能, 本研究首先利用 2.2 节设计的算法对如下的双峰概率密度函数进行采样研究.

$$p(\theta) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\theta^2\right] + \frac{2}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(2\theta - 8)^2\right] \quad (7)$$

该函数在 θ= 0 和 θ= 4 位置附近呈双峰特征.

3.2　WASP 模型应用(案例 2)

本案例以 WASP 模型系统在密云水库水质模拟中的应用为例来探讨 MCMC 法用于搜索水质模型参数后验分布及进行参数不确定分析的有效性和优越性. 有关模型时空概化、输入配置等内容详见文献[22].

WASP[23]是由美国国家环保局开发的用于地表水水质模拟的模型, 它提供了一个灵活的动态模拟系统. 如图 1 所示, WASP 可以模拟 8 个指标, 分别为: 氨氮(NH_3)、硝酸盐氮(NO_3)、溶解性磷酸盐(OPO_4)、叶绿素 a(Chl-a)、碳生化需氧量(CBOD)、溶解氧(DO)、有机氮(ON)和有机磷(OP). 在 WASP 模型系统中水质模块 EUTRO5 的水质参数有 42 个之多, 通过灵敏度分析提取了灵敏度较高的参数, 结果如表 1 所示.

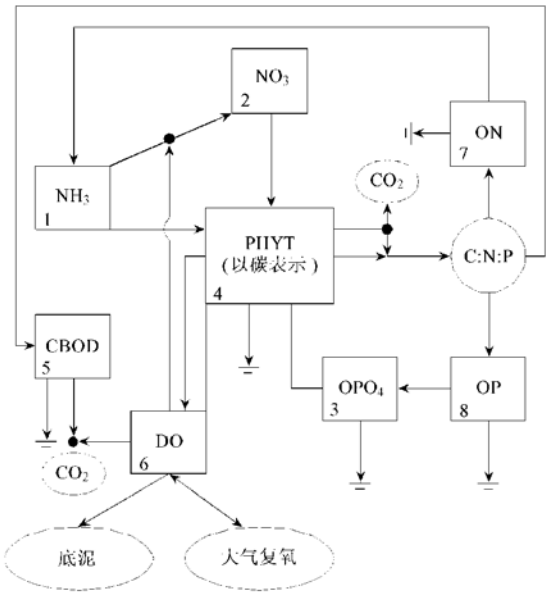

图 1　水质模拟反应动力学关系

Fig. 1　Wasp Eutrophication (EUTRO) state variables relationships

4　结果与讨论

4.1　案例 1

AM 算法配置为: 参数初始协方差矩阵 C_0 为对角矩阵, 取参数搜索范围的 1/20, 参数搜索范围为[-5, 10]; 初始迭代次数 i_0= 100. 算法平行运行 5 次, 每次采集1 000个样本, 摈弃前 200 个样本, 以消除初始化阶段的影响, 这样 5 次运行共4 000个样本, 图 2 为参数后验分布的直方图, 可以看出, 式(7)中参数 θ 的后验分布得到了很好地采样, 同时按式(6)计算 $\sqrt{R}$= 1.0001, 说明参数样本收敛到了后验分布.

表 1　待识别水质模型参数

Table 1　Parameters needed to identify in water quality model

参数名称	物理意义	参数范围
K12C	20 ℃条件下的硝化速度系数/d^{-1}	0.05~ 0.35
K20C	20 ℃条件下的反硝化速度系数/d^{-1}	0.01~ 0.2
K1C	浮游植物的饱和生长率/d^{-1}	1.5~ 4
K1RC	20 ℃条件下浮游植物的内源呼吸速率/d^{-1}	0.05~ 0.2
NCRB	浮游植物内的氮碳比(缺省值为 0.25) /mg•mg^{-1}	0.2~ 0.3
PCRB	浮游植物内的磷碳比(缺省值为 0.025) /mg•mg^{-1}	0.02~ 0.03
KDC	20 ℃ 条件下的 CBOD 降解速率/d^{-1}	0.01~ 0.1
K2	20 ℃ 条件下, 水体的复氧速度常数/d^{-1}	0.1~ 0.2
K71C	溶解有机氮的矿化速度/d^{-1}	0.01~ 0.1
K83C	溶解有机磷的矿化速度/d^{-1}	0.01~ 0.1

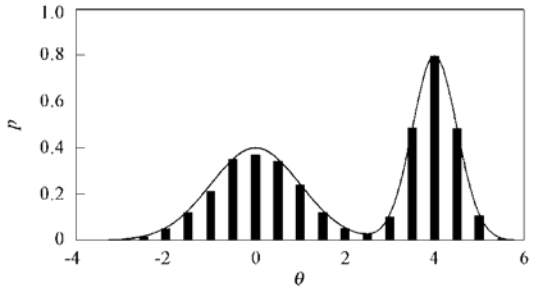

图 2　双峰概率密度函数和 MCMC 法采样 4 000 次的直方图

Fig. 2　Bimodal probability distribution and histogram of 4 000 samples generated using AM

4.2　案例 2

AM 算法控制参数设置如下: 参数初始协方差矩阵 C_0 为对角矩阵, 参数方差取参数搜索范围的 1/10, 参数搜索范围见表 1; 初始迭代次数 i_0= 1000. 利用 AM 算法搜索表 1 中参数的后验分布, 每次采样20 000次, 平行运行 5 次, 初始化阶段为10 000次(确定依据见后面分析), 这样 5 次平行试验共采集了50 000个样本用于参数后验分布的统计分析.

4.2.1　收敛性判断

采样序列是否已收敛到参数后验分布是后续研究的基础. 利用公式(6)计算的 $\sqrt{R}$ 演化过程如图 3 所示. 在搜索初期, 即迭代次数 i< 5000, 不同参数的 $\sqrt{R}$ 变化剧烈, 从 1.6~ 2.0 快速下降到 1.05 左右, i> 5000 后, $\sqrt{R}$ 继续缓慢下降并最终稳定到一个略大于 1.0 的数值上. 各参数均呈现出这一规律, 说明不同参数的 MCMC 采样序列均能稳定收敛到

参数的后验分布上. 由于每一序列均是随机搜索产生, 互不相关, 由公式(6) 可知, 若 $\sqrt{R}$ 接近 1, 说明各序列的均值和方差基本相同, 算法全局收敛.

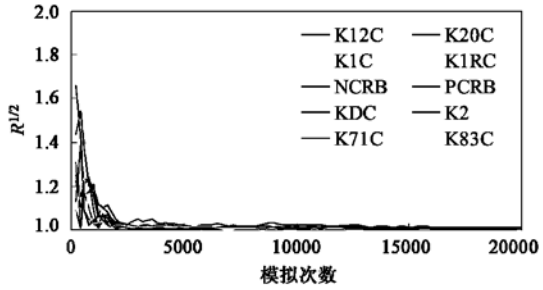

图 3 比例缩小得分 $R^{1/2}$演化过程

Fig. 3 Evolution of the scale reduction score

图 4 为以参数 K1C 为例的采样过程图, 可以看到, 参数取值遍历了整个参数空间. 图 5 和图 6 分别为参数 K1C 后验分布平均值和方差的变化过程图. 如图所示, 当 $i>$ 10000 后, 参数 K1C 的平均值和方差基本达到稳定. 由此可以得出, 每一序列是收敛的, 尽管不能确定它是全局收敛.

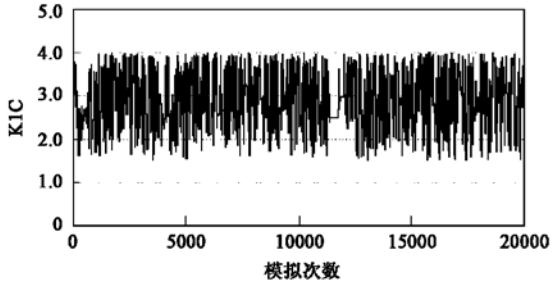

图 4 K1C采样过程

Fig. 4 Sampling trace of parameter K1C

4.2.2 参数后验分布

参数后验分布统计结果如表 2. 表 2 中分别统计了平均值、众数、标准差、斜度、最大值、最小值以及 7 个百分点数据, 其中众数不是通常意义上序列中出现概率最高的数, 而是将参数空间等分为 25 组, 取概率最高的一组的中值为众数. 百分点数据有助于确定置信区间, 进行参数的不确定性分析, 如 90% 置信度的参数区间为[5% 百分点, 95% 百分点], 80% 置信度的参数区间为[10% 百分点, 90% 百分点], 其它同理. 表 2 中的最小值和最大值与表 1 中的参数搜索范围基本一致, 说明参数空间得到有效搜索. 另外, 表 2 中的平均值和众数存在显著差别, 说明参数后验分布为非对称分布, 大多数参数呈偏正态分布, 斜度值也说明了这一点, 因此进行参数

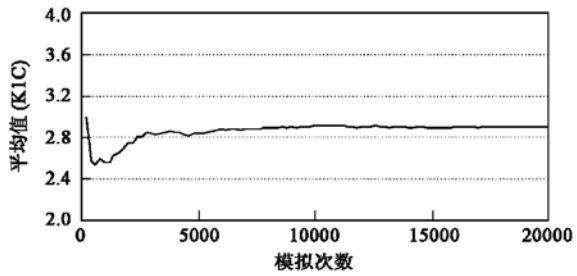

图 5 K1C后验分布平均值变化

Fig. 5 Posterior mean trace of parameter K1C

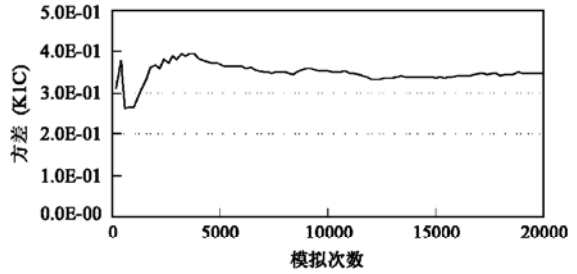

图 6 K1C后验分布方差变化

Fig. 6 Posterior variance trace of parameter K1C

表 2 参数后验分布

Table 2 Posterior statistics of Parameters

参数	样本数	平均值	众数	标准差	斜度	最小值	最大值	百分点						
								5%	10%	25%	50%	75%	90%	95%
K12C	50 000	0.218 7	0.200 0	0.076 2	−0.157 5	0.050 1	0.350 0	0.089 0	0.114 5	0.160 3	0.220 4	0.281 4	0.321 3	0.335 7
K20C	50 000	0.066 0	0.051 8	0.036 4	1.448 6	0.010 1	0.199 8	0.022 8	0.030 5	0.042 8	0.056 6	0.078 3	0.117 0	0.147 8
K1C	50 000	2.874 7	2.550 0	0.600 6	−0.047 1	1.501 0	3.999 9	1.873 2	2.077 1	2.430 3	2.861 5	3.363 7	3.705 1	3.845 6
K1RC	50 000	0.120 8	0.101 0	0.033 8	0.250 4	0.050 0	0.200 0	0.068 1	0.078 0	0.095 9	0.117 6	0.144 4	0.168 6	0.181 9
NCRB	50 000	0.249 1	0.234 0	0.028 8	0.035 6	0.200 0	0.300 0	0.204 8	0.209 5	0.224 2	0.248 3	0.274 0	0.289 3	0.294 4
PCRB	50 000	0.025 3	0.027 4	0.002 8	−0.121 6	0.020 0	0.030 0	0.020 6	0.021 2	0.022 8	0.025 4	0.027 7	0.029 0	0.029 5
KDC	50 000	0.042 2	0.037 0	0.022 5	0.860 3	0.010 0	0.100 0	0.014 1	0.017 9	0.025 6	0.035 5	0.054 9	0.079 4	0.089 1
K2	50 000	0.154 2	0.158 0	0.027 5	−0.176 2	0.100 0	0.200 0	0.107 8	0.114 4	0.131 8	0.156 4	0.177 8	0.190 6	0.194 8
K71C	50 000	0.045 7	0.026 2	0.022 5	0.689 7	0.010 0	0.100 0	0.017 3	0.021 2	0.028 1	0.040 1	0.061 0	0.081 7	0.090 7
K83C	50 000	0.038 7	0.033 4	0.015 5	1.541 3	0.010 0	0.100 0	0.021 0	0.024 1	0.029 0	0.034 7	0.043 8	0.059 7	0.073 2

不确定性分析时, 采用众数分析相对合理.

如前文分析, 优化算法的搜索结果为一组最优参数, 即目标函数达到或接近极小的参数值, 而 MCMC 采样算法的最终结果是参数的后验分布, 不再是一组参数值. 很容易理解, 优化算法的最优解应该位于参数后验分布的高概率密度区域, 为此, 研究比较了 MCMC 法不确定分析结果和遗传算法的搜索结果[24], 如前所述, 众数代表参数空间的高密度区域的中心, 因此表 3 给出了 MCMC 法的众数和优化算法搜索结果的平均值, 并计算了它们之间的相对误差. 由表 3 可知, 大多数参数的优化平均值和参数后验分布的众数之间的差别很小, 基本可以控制在 10% 的范围内, 个别参数误差较大的原因包括: ①参数灵敏度较低, 如 KDC、K71C; ②参数相关性, 如 NCRB 和 PCRB, 这 2 个参数的相关性从参数的物理意义便可清楚解释, NCRB 为浮游植物的氮碳比, PCRB 为浮游植物的磷碳比. 参数灵敏度和相关性的影响从参数后验分布直方图可以更清楚地看出. 可见当模型存在弱可识别性或不可识别参数时, 优化算法的搜索结果可能会失真.

图 7 为 AM 算法全局收敛情况下的部分参数后验分布直方图. 从图 7 (c) 可以看出, 由于参数相关性的影响, 参数 NCRB 基本呈均匀分布, PCRB 同样如此, 限于篇幅未给出. 参数灵敏度的影响在参数后验分布中也可反应出来, 如图 7 (d) 中参数 KDC 的后验分布, 低灵敏度导致参数的不确定性增大. 相比之下, 高灵敏度参数, 如 K20C 和 K1C, 无论优化算法, 还是 MCMC 法, 均可有效识别, 见图 7 (a) ~ (b).

可见, MCMC 法能有效搜索参数后验分布, 适合于复杂模型的参数识别和不确定性分析.

表 3　MCMC 法和遗传算法结果对比

Table 3　Comparison between results of MCMC and Genetic Algorithm

参数名称	MCMC 法众数	遗传算法优化值	相对误差/%
K12C	0.200 0	0.205 3	2.6
K20C	0.051 8	0.050 6	2.5
K1C	2.550 0	2.516 3	1.3
K1RC	0.101 0	0.100 6	0.4
NCRB	0.214 0	0.249 7	14.3
PCRB	0.027 4	0.024 6	11.6
KDC	0.037 0	0.030 4	21.6
K2	0.158 0	0.151 3	4.4
K71C	0.026 2	0.030 4	13.7
K83C	0.031 4	0.029 8	5.5

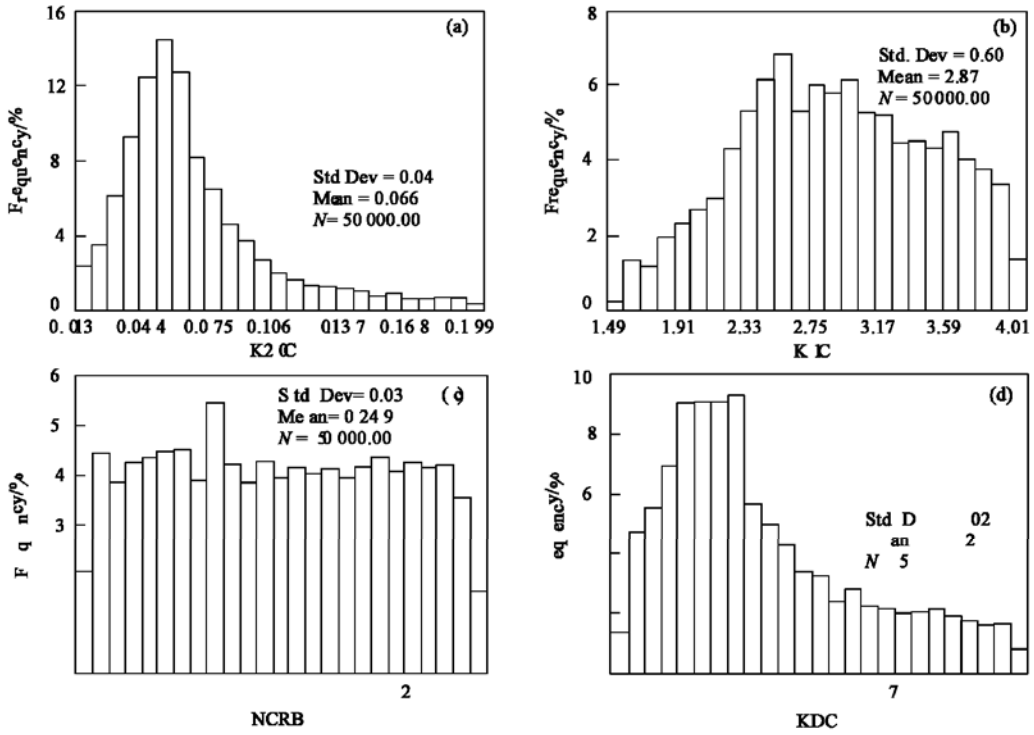

图 7　参数后验分布直方图

Fig. 7　Histogram of parameter posterior distribution

5 结论

(1) 研究采用自适应 Metropolis 算法来搜索参数后验分布. 为考察 MCMC 法对参数后验分布的搜索性能和效率, 进行了 2 个案例研究. 案例 1 为一个已知的双峰概率密度函数, 案例 2 为 WASP 模型在密云水库水质模拟的应用研究. 结果表明, MCMC 法能有效搜索参数后验分布, 适合于复杂模型的参数识别和不确定性分析. 同时, Gelman 收敛判别准则计算表明, MCMC 采样序列均能稳定收敛到参数的后验分布上.

(2) 相比基于随机采样的方法, 如 RSA 法和 GLUE 法, MCMC 法对参数后验分布的搜索, 无论是搜索性能还是搜索效率, 均表现出了独特的优越性, 同时, AM 算法可有效处理参数灵敏度和参数相关性的影响.

参考文献:

[1] 程声通, 陈毓龄. 环境系统分析[M]. 北京: 高等教育出版社, 1990.

[2] Chen J, Wheater H S. Identification and uncertainty analysis of soil water retention models using lysimeter data[J]. Water Resources Research, 1999, **35**(8): 2401~ 2414.

[3] 刘毅, 陈吉宁, 杜鹏飞. 环境模型参数识别与不确定性分析[J]. 环境科学, 2002, **23**(6): 6~ 11.

[4] Beck M B. Water Quality Modeling: A Review of the Analysis of Uncertainty[J]. Water Resources Research, 1987, **23**(8): 1393~ 1442.

[5] Omlin M, Reichert P. A comparison of techniques for the estimation of model prediction uncertainty[J]. Ecological Modelling, 1999, **115**: 45~ 59.

[6] Beven K. Binley A. The future of distributed models: model calibration and uncertainty prediction[J]. Hydrological processes, 1992, **6**: 279~ 298.

[7] Reichert P, Omlin M. On the usefulness of overparameterized ecological models[J]. Ecological Modelling, 1997, **95**: 289~ 299.

[8] Tiwari J, Hobbie J, Peterson B. Random differential equations as models of ecosystems, Ⅲ Bayesian inference for parameters[J]. Math. Biosci., 1978, **38**: 247~ 258.

[9] Hornberger G M, Spear P C. Eutrophication in Peel Inlet-Ⅰ. The problem-defining behaviour and mathematical model for the phosphorus scenario[J]. Wat. Res., 1980, **14**: 29~ 42.

[10] Spear R C, Hornberger G M. Eutrophication in Peel Inlet-Ⅱ. Identification of critical uncertainties via generalized sensitivity analysis[J]. Wat. Res., 1980, **14**: 43~ 49.

[11] 邓义祥, 王琦, 赖斯芸, 等. 优化 RSA 和 GLUE 方法在非线性环境模型参数识别中的比较[J]. 环境科学, 2003, **24**(6): 9~ 15.

[12] 茆诗松. 贝叶斯统计[M]. 北京: 中国统计出版社, 1999.

[13] Jorgensen S E. An improved parameter estimation procedure in lake modeling[J]. Lake & Reservoirs: Research and management, 1998, **3**: 139~ 142.

[14] 龚光鲁, 钱敏平. 应用随机过程教程及其在算法与智能计算中的应用[M]. 北京: 清华大学出版社, 2003.

[15] Gilks W R, Richardson S, Spiegelhalter D J. Markov chain Monte Carlo in practice[M]. London: Chapman & Hall, 1996.

[16] Smith A F M, Robert G O. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods[J]. Journal of Royal Statistical Society Series B, 1993, **55**: 3~ 23.

[17] Tierney L. Markov-chains for exploring posterior distributions[J]. Annals of Statistics, 1994, **22**: 1701~ 1762.

[18] Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm[J]. American Statistician, 1995, **49**(4): 327~ 335.

[19] Haario H, Saksman E, Tamminem J. An adaptive Metropolis algorithm[J]. Bernoulli, 2001, **7**(2): 223~ 242.

[20] Roberts G O, Tweedie R L. geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms[J]. Biometrika, 1996, **83**: 95~ 110.

[21] Gelman A, Rubin D B. Inference from iterative simulation using multiple sequences[J]. Statistics Science, 1992, **7**(4): 457~ 511.

[22] 贾海峰. GIS 强化下的水库水质模拟及其在密云水库中的应用研究[D]. 北京: 清华大学, 1999.

[23] Ambrose R B, Wool T A, Martin J L, *et al*. WASP5.x, A Hydrodynamic and Water Quality Model Model Theory, User's Manual, and Programmer's Guide[M]. Draft: Environmental Research Laboratory, US Environmental Protection Agency, 1993.

[24] 王建平, 程声通, 贾海峰. 水质模型参数优化的遗传算法实现及控制参数分析[J]. 环境科学, 2005, **26**(3): 61~ 65.