文章编号:1000-6788(2008)01-0100-09

Metropolis-Hastings 自适应算法及其应用

陈　平[1],徐若曦[2]

(1. 东南大学 数学系,南京 210096;2. 俄亥俄州立大学 统计系,美国;43210-1247)

摘要: 首先阐述 Metropolis-Hastings 算法实现的具体步骤,然后证明由此产生的 Markov 链满足细致平衡条件,从而以目标分布为不变分布.接下来给出几个计算实例,以说明提议函数及其方差的选取对采样结果的影响,并由此推出一种改进的自适应算法用以寻找合适的提议函数及其方差.最后,通过贝叶斯 Logistic 模型的例子说明 M-H 方法在贝叶斯分析中的应用,同时也检验 M-H 自适应算法的效果.

关键词: MCMC;Metropolis-Hastings 算法;马尔可夫链;R 软件;贝叶斯分析

中图分类号: F830;TP183　　**文献标志码:** A

Metropolis-hastings adaptive algorithm and its application

CHEN Ping[1], XU Ruo-xi[2]

(1. Department of Mathematics, Southeast University, Nanjing 210096, China; 2. Department of Statistics, The Ohio State University, Columbus, OH 43210-1247)

Abstract: Markov chain Monte Carlo (MCMC) methods is an important class of computer based simulation techniques. This paper investigates one MCMC method known as the Metropolis-Hastings algorithm. In this paper, we first introduce readers the proceedings of the Metropolis-Hastings algorithm. Then we prove the resulting chain satisfies detailed balance, and hence has the target distribution as the invariant distribution. Next, we provide some illustrative examples that show the influence of the proposal function and its variance on the resulting chain, and develop an adaptive method to find optimal proposal for the random walk sampler. Finally, we discuss the relationship between M-H algorithm and Bayesian analysis. The Bayesian Logistic model is used to illustrative the application of M-H algorithm in Bayesian analysis and to test the proposed adaptive method.

Key words: MCMC; Metropolis-Hastings algorithm; markov chain; R software; bayesian analysis

1　引言

MCMC(Markov chain Monte Carlo)方法又被称为动态 Monte Carlo 方法,它是以动态构造 Markov 链为基础,通过遍历性约束来实现模拟目标分布的一类随机模拟方法.在过去的 20 年里,MCMC 方法对统计学,尤其是贝叶斯分析的发展产生了深远的影响.本文借助于 R 软件对 MCMC 方法中的 Metropolis-Hastings(M-H)算法作进一步的探讨,并由此推出一种改进的自适应 M-H 算法,从而可以提高算法的效率.同时还将借助于贝叶斯分析的实例阐明 M-H 算法与现代统计推断间的紧密联系.

假设我们要计算函数 $F(x)$关于概率密度 $\pi(x)$的平均值,即

$$E[F(x)] = \int_S F(x)\pi(x)\mathrm{d}x$$

如果这个积分用分析的方法无法解决,那么随机模拟的方法就可以用来估计积分值.前提是我们能够从 $\pi(x)$中产生独立同分布的随机数.如果$(X_1,\cdots,X_n)$i.i.d. $\sim \pi(x)$,那么依据大数定律,$\frac{1}{n}\sum_i^n F(X_i)$依概率收敛到 $\int_S F(x)\pi(x)\mathrm{d}x$.于是问题转化为如何从概率分布 $\pi(x)$中产生独立同分布的随机数.虽然在很多

收稿日期:2006-07-13

资助项目:国家自然科学基金(10671032)

作者简介:陈平(1960-),男,江苏溧阳人,教授,博士,研究方向:金融与工程时间序列分析、生存分析等,E-mail:cp18@263.net.cn.

万方数据

问题中,直接产生独立同分布的随机数并不容易,但是很多 MCMC 方法,如 Gibbs 采样法,Metropolis 采样法,却能够通过随机模拟产生收敛到 $\pi(x)$的 Markov 链.例如,通过 100 次模拟产生 100 条长度为 5000 的独立的 Markov 链,那么每条链的最后一个值 X_{5000}就可以近似地看作来自极限分布 $\pi(x)$的大小为 100 的独立同分布的样本.当然,在实际计算中,更高效的做法是在 Markov 链运行了一段时间(比如 500 次)后,间隔 k 个点采样一次,得到的样本仍可近似看作来自 $\pi(x)$的独立同分布的样本.本文主要讨论 MCMC 方法中很重要的一类算法:M-H 算法.我们将分析 M-H 算法构造 Markov 链的方法,证明其具有不变分布(极限分布),探讨影响算法效率的因素,以及算法在贝叶斯分析中的一些应用.

具体安排是:第二节阐述 Metropolis-Hastings 算法的实现过程及一些常用的形式,给出算法可逆性的证明;第三节给出一些计算实例,提出 M-H 自适应算法及改进一般算法的途径并通过一些例子做比较分析.第四节给出贝叶斯分析的有关理论,通过一个具体的贝叶斯模型例子来阐明 M-H 自适应算法在贝叶斯分析中的应用,同时再次检验第三节中提出的算法的效果.最后给出简要结论.

2　Metropolis-Hastings 算法及其性质

M-H 算法的思想是构造一个以目标分布 $\pi(x)$为不变分布的 Markov 链.为了实现这一目标,M-H 算法借助于一个辅助的概率密度函数 $q(x,y)$,通常被称为"提议函数".$q(x,y)$通常要满足以下三个条件(其中 S 表示状态空间):(a)对于固定的 x,$q(x,\cdot)$是一个概率密度函数.(b)对于 $\forall x,y\in S$,$q(x,y)$的值要能够计算出来.(c)对于固定的 x,能够方便地从 $q(x,y)$中产生随机数.

在满足上述三个条件的情况下,$q(x,y)$的选取是任意的.下面是算法的具体步骤,假设当前状态 $X_n=x$,那么

1) 从提议函数 $q(x,\cdot)$中产生一个新状态 y.

2) 计算接受概率

$$\alpha(x,y)=\min\left\{1,\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} \tag{1}$$

3) 以概率 $\alpha(x,y)$置 $X_{n+1}=y$;以概率 $1-\alpha(x,y)$置 $X_{n+1}=x$.

我们将证明此算法构造的 Markov 链以目标分布 π 为不变分布.

从理论上讲,提议函数 $q(x,y)$的选取是任意的,但在实际计算中,提议函数的选取对于算法的效率的影响是相当大的.一般认为提议函数的形式与目标分布越接近,则模拟的效果越好.我们将在第三节里仔细讨论提议函数的选取问题.在算法的第 3 步中,如果置 $X_{n+1}=y$,我们就称之为"接受提议";否则就称为"拒绝提议".在实际计算中以概率 $\alpha(x,y)$接受 y 可以这样实现:从 $U(0,1)$产生一个随机数 u,如果 $u<\alpha(x,y)$,则置 $X_{n+1}=y$;否则置 $X_{n+1}=x$.如果将一个使得 $\pi(x)\neq 0$ 的点 x 作为整条 Markov 链的起始点,那么(1)式中比值的分母将不会为零.原因是使得 $q(x,y)=0$ 的新状态 y 被提出的概率为零,同时如果 $\pi(y)=0$,那么 y 被接受的概率为零,提议将被拒绝.所以只要 $\pi(X_1)>0$,那么在后续的计算中分母为零的概率将是零,从而不会对实际计算造成麻烦.

如果 M-H 算法中的提议函数 $q(x,y)$不仅满足对称性,而且只与 $x-y$ 有关,那么算法就演变为通常意义下的随机游动采样法.最常见的一种随机游动采样法以正态分布,即 $q(x,y)=\phi(x-y)$为提议函数.这里 ϕ 是任意一种多维正态分布的密度函数,也就是说 $y\sim N(x,\Sigma)$.(Σ 是任意的正定矩阵,x 是当前的状态).在实际计算中,由于 Σ 要求是正定矩阵,所以常取为 $\Sigma=\sigma I$,σ 是一个参数.这时一个重要的问题是 σ 应取多少才合适.如果取的过大,那么大多数提出的新状态将落入分布的尾部,从而被拒绝.如果取的过小,那么采样的自相关系数就会很高,而且收敛到目标分布的速度也会很慢.不论哪种情况,我们都不能在时间有限的采样过程中得到较好地服从目标分布的随机数.一般认为在模拟多维正态随机数时,如果调整 σ 的值使得在整个模拟过程中提议被接受的比例大约为 20%,将得到比较好的模拟效果.

在时齐且有条件转移密度的情形,我们简记

$$P(x,y)\equiv P^{(n,n+1)}(x,y)=P^{(0,1)}(x,y)$$

本文所考虑的 Markov 链都是时齐的,并总假定条件转移密度存在.M-H 算法构造的 Markov 链的条件转移

万方数据

密度为:

$$P(x,y)=\begin{cases}q(x,y)\alpha(x,y), & \forall y\neq x\\ 1-\int q(x,y)\alpha(x,y)\mathrm{d}y, & y=x\end{cases} \tag{2}$$

下面的定理说明 M-H 算法构造的 Markov 链以目标分布 $\pi(x)$为不变分布,并且是时间可逆的.

定理 1 若以 P 为条件转移密度的马尔科夫链具有可逆初分布π,那么 π 也一定是其不变分布.

证明

$$Pr(X_{n+1}=y)=\int\pi(x)P(x,y)\mathrm{d}x=\int\pi(y)P(y,x)\mathrm{d}x$$
$$=\pi(y)\int P(y,x)\mathrm{d}x=\pi(y)$$

y 的边缘分布即为π,这就证明了这个定理.

有了这个定理,我们下面所要做的就是验证 M-H 算法构造的 Markov 链满足细致平衡条件.

定理 2 设 $P(x,y)$是由(2)式定义的条件转移密度.则在状态空间 S 上,对于任意的初始分布 π,以 P 为条件转移密度的 Markov 链满足细致平衡条件(时间可逆的),即

$$\pi(x)P(x,y)=\pi(y)P(y,x),\quad \forall x,y\in S$$

证明 如果 $y=x$,那么自然 $\pi(x)P(x,x)$与其自身相等.对于 $y\neq x$ 的情况,

(ⅰ)若

$$\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\leqslant 1$$

则

$$\alpha(x,y)=\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\quad 并且\quad \alpha(y,x)=1$$

从而有

$$\pi(x)P(x,y)=\pi(x)q(x,y)\alpha(x,y)=\pi(x)q(x,y)\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$$
$$=\pi(y)q(y,x)=\pi(y)q(y,x)\alpha(y,x)=\pi(y)P(y,x)$$

(ⅱ)若

$$\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\geqslant 1$$

则

$$\alpha(y,x)=\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\quad 并且\quad \alpha(x,y)=1$$

从而有

$$\pi(y)P(y,x)=\pi(y)q(y,x)\alpha(y,x)=\pi(y)q(y,x)\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}$$
$$=\pi(x)q(x,y)=\pi(x)q(x,y)\alpha(x,y)=\pi(x)P(x,y)$$

不论哪种情况,均有 $\pi(x)P(x,y)=\pi(y)P(y,x)$,证毕.

3 两种 Metropolis-Hastings 算法比较

R 软件是一种基于 S 语言而开发的用于数据处理,数值计算以及图象显示的免费软件.虽然 R 软件不仅仅用于统计计算,但多数人将其当作一种统计软件来使用,原因是 R 软件的工作环境中集成了很多经典或现代的统计计算方法.其中一部分包含在基本工作环境中,更多的是以数据包的形式提供给使用者.数据包是 R 软件集成环境的体现,任何使用者都可以编写数据包,供自己或他人使用.R 软件的基本工作环境自带约 25 个标准数据包,更多其他的数据包可以从任一 CRAN(Comprehensive R Archive Network)站点及其镜象站点获得.

万方数据

例 1　假设目标分布函数为 $N(3,5)$，提议函数采用随机游动，即新状态 $y \sim N(X_n,\sigma)$，其中 X_n 为当前状态，σ 为标准差，取为 2. M-H 算法的步骤如下：

1) 取初值 $X_1 = 100$.

2) 从提议函数 $N(X_n,\sigma)$ 中产生一个新状态 y.

3) 计算接受概率

$$\alpha(x,y) = \min\left\{1,\frac{\pi(y)}{\pi(x)}\right\} = \min\left\{1,\exp\left(-\frac{1}{50}(y^2 - x^2 - 6y + 6x)\right)\right\}$$

其中 $\pi(x)$ 为正态分布 $N(3,5)$ 的密度函数.

4) 以概率 $\alpha(x,y)$ 置 $X_{n+1} = y$；以概率 $1-\alpha(x,y)$ 置 $X_{n+1} = x$.

我们构造了一条长度为 10000 的 Markov 链，为了消除初值对采样的影响，从第 500 个样本开始取样分析. 从直方图上看采样结果令人满意，估计的密度函数与目标分布函数十分接近. 但为了得到近似独立的样本，往往采用间隔采样法. 从样本的自相关系数图(略)中可看出在 lag = 40 左右时，样本的自相关系数已接近于零，故若我们每隔 40 个点取样一次，则取出的样本可近似看作独立的样本. 但这样一来，我们从这条链中最多只能取得 250 个近似独立的样本. 因此，在实际工作中，总是希望 lag 能取的越小越好.

进一步用不同的 σ 的值($\sigma = 1,2,3,10,15,40,100,200$)重复上面的实验，得到的结果显示：当取 $\sigma = 10$ 或 15 时，用同样的算法得到的 Markov 链可以用不到 20 的间隔进行采样，从而所能得到的近似独立的样本大小将两倍于前者，效率大大提高. 这就是说 σ 的选取确实对算法有很大的影响，但是 σ 的选取和哪些因素有关呢？是否存在普遍适用的值呢？为此，我们再用一个例子来说明这一点.

例 2　我们将目标分布换为 Beta(3,5)，提议函数仍然采用随机游动，即新状态 $y \sim N(X_n,\sigma)$，其中 X_n 为当前状态. 同样，我们对不同的 σ 值($\sigma = 0.05,0.1,0.4,0.6,1,3,10,15$)进行重复实验，得到的结果如图 1 所示.

从图 1 中易见较为合适的 σ 值为 0.4,0.6,1. 若如上例一样取 10 或 15 作为 σ 的值，则无法得到较好的采样结果. 例 1 与例 2 说明 σ 的选取要考虑目标分布. 对应于不同的目标分布，合适的 σ 值可能相差甚远，在实际计算中必须具体问题具体分析.

随机游动采样法是 M-H 算法中较为常用的一种形式，我们已经看到对于随机游动采样法而言，如何选择提议函数的方差成为影响算法效率的主要因素. 对于一维情况，就是要选择一个合适的 σ 值，而对于 $d > 1$ 的多维情况，我们需要一个合适的正定方差矩阵 Σ. 那么现在的问题是什么样的 σ 值或者方差矩阵 Σ 才是合适的呢？一种想法是调整接受概率到合适的值. 根据关于采用多维正态作为提议函数的随机游动采样法的一些理论结果. 一般认为如果目标分布是 d 维的且各分量之间独立，则调整接受概率到 0.234 左右将最大限度地优化随机游动采样法的效率，并且接受概率 0.234 与目标分布的具体形式无关.

基于这样的想法，我们尝试构造一种自适应算法来寻找合适的提议函数的方差. 大致的思想是让接受比率落入一个可以接受的范围内，比如区间 $[a-\varepsilon, a+\varepsilon]$. 以一维的情况为例，先给定一个初值 σ_0，用这个值产生一条链，来估计接受比率 P_0. 若 P_0 在给定的区间中，则已符合要求，算法停止；若不然，则以 σ_0 为中心做一次随机游动，得到一个 σ_1，再计算 P_1 是否满足要求，以此类推，直至找到合适的 σ 值. 具体的实现步骤如下：

1) 给定 σ 一个初值 σ_0，置 $n = 0$.

2) 以 $N(X_{i-1},\sigma_n)$ 为提议函数，用随机游动采样法产生一条长度为 m 的 Markov 链.

3) 计算第 2 步中产生的 Markov 链接受新状态的比率 P_n.

4) 如果 $P_n \in [a-\varepsilon, a+\varepsilon]$，则置 $\sigma = \sigma_n$，退出；如果 $P_n \notin [a-\varepsilon, a+\varepsilon]$，转入第 5 步.

5) 如果 $n > 2$ 并且 $|P_n - a| < |P_{n-1} - a|$，则置 $\sigma_n = \sigma_{n-1}$.

6) 从 $N(\sigma_n,\hat{\sigma})$ 中产生一个随机数，记为 σ_{n+1}，置 $n = n+1$，转入第 2 步.

注　第 5 步的判断是为了保证接受比率 P_n 要优于 P_{n-1}，即 σ_n 至少不会比 σ_{n-1} 要坏；第 6 步中的 $\hat{\sigma}$ 的选取没有固定的规律可循，主要是看对初始 σ_0 的估计是否接近于合适的 σ 值. 若估计 σ_0 与 σ 较为接近，

万方数据

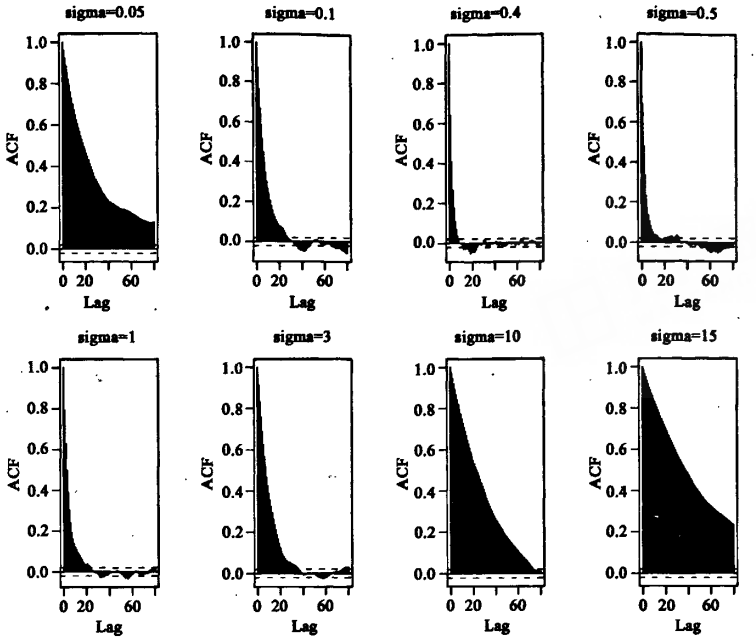

图 1 取 $\sigma = 0.05, 0.1, 0.4, 0.6, 1, 3, 10, 15$ 时，随机游动采样法得到的样本自相关系数图．横坐标为滞后系数 lag，纵坐标为对应的自相关系数，两条虚线表示自相关系数为零的 95% 的置信区间．

则 $\hat{\sigma}$ 可选择得小一些；若估计 σ_0 与 σ 相差较远，则 $\hat{\sigma}$ 可选择得大一些；对于 $d > 1$ 的多维情况，如果要确定一个任意的正定矩阵 Σ，则要确定 d^2 个参量，显得有些麻烦．更多情形下，我们取 $\Sigma = \sigma I$．这样一来要确定的参数就只有一个，同时上述算法也可以用于多维情形，只须在 6 步中“以 $N(X_{i-1}, \sigma_n)$ 为提议函数”改为“以 $N(X_{i-1}, \sigma_n I)$ 为提议函数”即可．

例 3 用例 2 的 Beta(3,5)分布来检验上述自适应算法．取 $\sigma_0 = 10, m = 2500, a = 0.23, \varepsilon = 0.01, \hat{\sigma} = 1$，计算得到符合要求的 σ 值为 0.82．图 2 为采样结果对比图，第一行为未经自适应算法改进得到的样本的自相关系数图，轨迹图及依据样本估计的密度曲线；第二行为经过自适应算法改进得到的样本的自相关系数图，轨迹图及依据样本估计的密度曲线．

从第一列自相关系数图的比较可看出经过自适应算法改进得到的样本要更符合实际计算需要，这表现为其自相关系数图为一个很陡的下降曲线，并且在 lag = 20 左右其自相关系数就已可以视为零了．相比较而言，未经自适应算法改进得到的样本的自相关系数图为一个较平缓的下降曲线，且在 lag = 80 时自相关系数仍然很大，以致不能视为零．第二列两种算法轨迹图的比较显示经过自适应算法改进得到的 Markov 链移动的更快，从而能够更有效的搜索整个状态空间．从第三列估计的密度曲线图可明显看出，依据经过自适应算法改进得到的样本估计出的密度曲线图要比依据未经过自适应算法改进得到的样本估计出的密度曲线图要光滑．

4 Metropolis-Hastings 算法与贝叶斯分析

贝叶斯估计方法就是把未知参数 θ 视为一个具有已知分布 $\pi(x)$ 的随机变量，从而将先验信息数字形式化并加以利用的一种方法，通常称 $\pi(x)$ 为先验分布．运用 M-H 方法，我们可以从复杂的多维分布中采样，从而能够对贝叶斯估计方法进行很好的模拟计算．

设总体 $\mathscr{X}$ 的分布密度为 $p(x, \theta), \theta \in \Theta$，$\theta$ 的先验分布为 $\pi(\theta)$，由于 θ 为随机变量并假定已知 θ 的先验分布，所以总体 $\mathscr{X}$ 的分布密度 $p(x, \theta)$ 应看作给定 θ 时 $\mathscr{X}$ 的条件分布密度，于是总体 $\mathscr{X}$ 的分布密度 $p(x, \theta)$ 需改用 $p(x|\theta)$ 来表示．设 $X = (X_1, \cdots, X_n)$ 为取自总体 $\mathscr{X}$ 的一个样本，当给定样本值 $x = (x_1, \cdots,$

万方数据

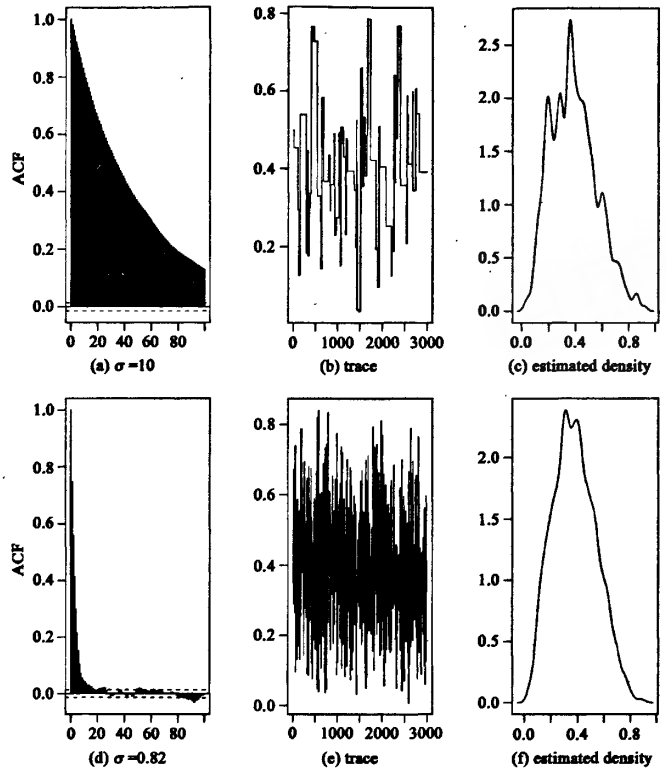

图 2　图(a)~(c)分别为采用一般算法($\sigma=10$)进行采样得到的样本的自相关系数图,轨迹图及依据样本估计的密度曲线.图(d)~(f)则分别为采用自适应算法($\sigma=0.82$)进行采样得到的样本的自相关系数图,轨迹图及依据样本估计的密度曲线.

x_n)时,样本 $X=(X_1,\cdots,X_n)$的联合密度为:$q(x_1,\cdots,x_n\mid\theta)=\prod_{i=1}^{n}p(x_i\mid\theta)$,或表示为:$q(x\mid\theta)=\prod_{i=1}^{n}p(x_i\mid\theta)$,从而,样本 X 和θ 的联合分布密度为:$f(x,\theta)=q(x\mid\theta)\pi(\theta)$.由乘法公式知:

$$f(x,\theta)=\pi(\theta)q(x\mid\theta)=g(x)h(\theta\mid x)$$

于是有:

$$h(\theta\mid x)=\frac{\pi(\theta)q(x\mid\theta)}{g(x)},\quad \theta\in\Theta \tag{3}$$

它是给定样本后 θ 的条件分布.其中 $g(x)$是(X,θ)关于样本 X 的边缘分布.如果 θ 是连续型随机变量,则有:$g(x)=\int_{\theta}q(x\mid\theta)\pi(\theta)\mathrm{d}\theta$;如果 θ 是离散型随机变量,则有:$g(x)=\Sigma_{\theta}q(x\mid\theta)\pi(\theta)$.

贝叶斯估计方法认为后验分布集中体现了样本和先验分布两者所提供的关于总体的信息,因而应在后验分布的基础上来进行统计分析.在过去的实际应用中,由于 $g(x)$不便于用分析的方法计算,使得贝叶斯估计方法的应用大为受阻.为了方便计算而过度简化的模型常被用来凑合一下.M-H 方法的运用改变了这一状况.

由于在贝叶斯估计方法中 $g(x)$不依赖于 θ,在计算后验分布中仅起到一个正则化因子的作用,若把 $g(x)$省略,可将贝叶斯公式改写为如下等价形式:

$$h(\theta\mid x)\propto\pi(\theta)q(x\mid\theta) \tag{4}$$

其中符号"$\propto$"表示两边仅相差一个不依赖于 θ 的常数因子.(4)式的右端虽不是正常的密度函数,但它是后验分布 $h(\theta\mid x)$的主要部分.不难看出,M-H 算法用于此类模型的一个主要优势在于只要(4)式右边已知,即可从后验分布中采样,从而避免了计算 $g(x)$的问题.

万方数据

在贝叶斯 Logistic 回归中假定样本观测值 $Y=(y_1,y_2,\cdots,y_n)$服从参数为 p 的两点分布，即

$$P(y_i=1)=p(\theta_i),\quad i=1,2,\cdots,n$$

其中

$$p(\theta_i)=logit^{-1}(\theta_i)=\frac{e^{\theta_i}}{1+e^{\theta_i}},\quad i=1,2,\cdots,n \tag{5}$$

并且

$$(\theta_1,\theta_2,\cdots,\theta_n)^{\mathrm{T}}=X\beta=\begin{pmatrix}\beta_0+x_{11}\beta_1+\cdots+x_{1k}\beta_k\\ \beta_0+x_{21}\beta_1+\cdots+x_{2k}\beta_k\\ \cdots\cdots\\ \beta_0+x_{n1}\beta_1+\cdots+x_{nk}\beta_k\end{pmatrix}.$$

变换函数 $logit(p)=\log[p/(1-p)]$，$logit^{-1}$为 $logit$ 的反变换，X 为 $n\times k$ 的矩阵，β 为 $k\times 1$ 的列向量．y_i，$i=1,2,\cdots,n$ 为取值为零或一的随机样本，x_{ij}为任意的实数．

例 4 失效的密封圈

1986 年，“挑战者”号航天飞机在发射升空 72 秒后爆炸，造成 7 名宇航员罹难，事故原因怀疑是密封圈失效．下表为航天飞机在不同温度下发射时密封圈失效的记录．

表 1 航天飞机在不同温度下发射时密封圈失效的记录

飞行号	14	9	23	10	1	5	13	15	4	3	8	17
失效(=1)	1	1	1	1	0	0	0	0	0	0	0	0
温度(华氏)	53	57	58	63	66	67	67	67	68	69	70	70
飞行号	2	11	6	7	16	21	19	22	12	20	18	
失效(=1)	1	1	0	0	0	1	0	0	0	0	0	
温度(华氏)	70	70	72	73	75	75	76	76	78	79	81	

我们用贝叶斯 Logistic 回归模型对表 1 中的数据做分析．设 x 表示发射时的温度，$p(\theta)$表示在不同温度时密封圈失效的概率，$y=1$ 表示密封圈失效，$y=0$ 表示密封圈没有失效．

现在已有一组观测值(x_i,y_i)，$i=1,2,\cdots,23$，按照模型假设

$$y_i\sim B(1,p(\theta_i))\quad 并且\quad p(\theta_i)=\frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}}$$

先验分布取为：$\pi(\beta)\equiv 1$，似然函数为：

$$L(\beta\mid Y)\propto\prod_{i=1}^{23}\left(\frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+x_i\beta_1}}\right)^{1-y_i} \tag{6}$$

后验分布正比于似然函数．

我们先用类似于例 1 的随机游动采样法进行模拟，取提议函数为 $N(X_{i-1},1)$，构造一条长度为 10000 的 Markov 链．模拟结果见图 3．

图中显示估计 β_0 的 Markov 链几乎没有平稳性可言．同时 0.95% 的接受比率也使得序列的自相关系数非常高，因而，模拟结果很难令人满意．观察发现回归系数 β_0 与 β_1 高度相关．这启发我们如果对回归变量进行变换，可能会得到较好的结果，比如用 $x-\bar{x}$ 作为新的回归变量．设 β'_0，β'_1 为变换后的回归系数，则由：$\beta_0+x\beta_1=\beta'_0+(x-\bar{x})\beta'_1$ 可得：

$$\beta_0=\beta'_0-\bar{x}\beta'_1,\beta_1=\beta'_1 \tag{7}$$

有了(7)式，我们就可以方便地转回到初始的回归系数值上．经过变换后，模拟的结果有了很大的改进．从图象上看，β_0，β_1 的轨迹

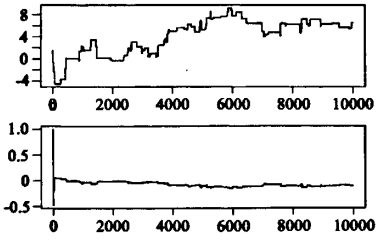

图 3 上图为 β_0 的轨迹图，下图为 β_1 的轨迹图

万方数据

都已表现出平稳的态势,其自相关系数也缩小到可以接受的范围以内.进一步研究发现,经过变换后,以 $N(X_{i-1},1)$ 为提议函数进行模拟计算已经能够得到较为满意的结果,但在上面的模拟过程中,接受的比率大约为 10%,仍然偏低.为了进一步改进得到的样本的性质,我们用第 3 节中提到的自适应算法对这一问题做模拟计算,从计算得到的 β_0,β_1 的轨迹对比图(图 4)中易见自适应算法得到的样本轨迹比一般随机游动算法得到的样本轨迹移动的更加迅速,也更加的平稳.自相关系数对比图(图 5)则显示,不论是对 β_0 还是 β_1 而言,自适应算法得到的样本的自相关系数都比一般随机游动算法得到的要小.这说明经过改进的自适应算法更具有优势.

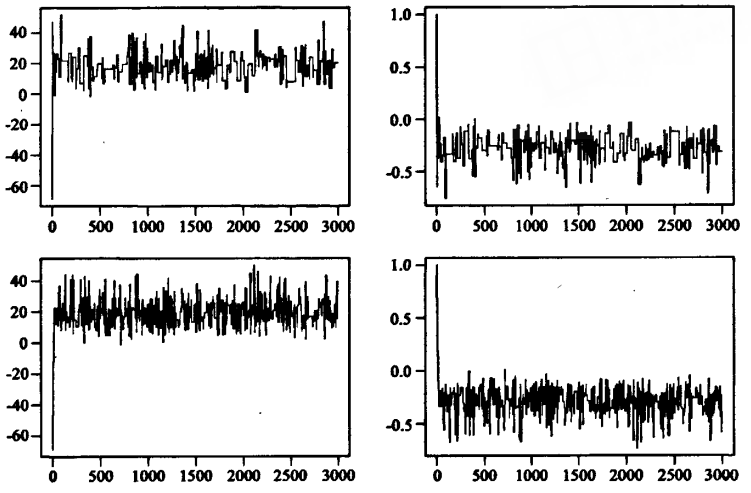

图 4　一般随机游动与自适应算法得到的样本的轨迹对比图.第一行为一般随机游动算法得到的 β_0,β_1 的轨迹,第二行为自适应算法得到的 β_0,β_1 的轨迹

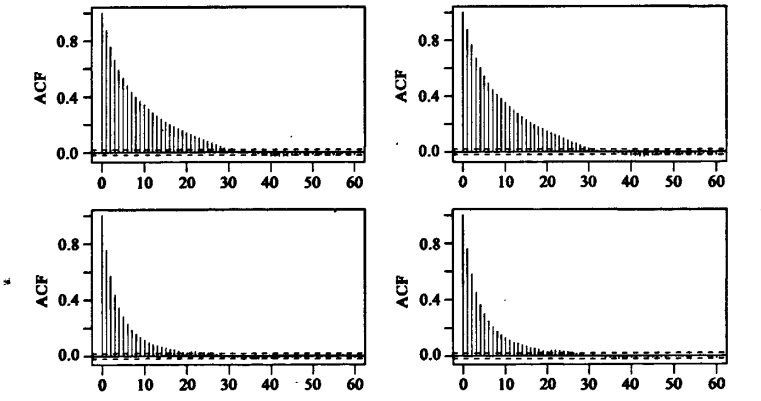

图 5　一般随机游动与自适应算法得到的样本的自相关系数对比图.第一行为一般随机游动算法得到的 β_0,β_1 的自相关系数图,第二行为自适应算法得到的 β_0,β_1 的自相关系数图

值的一提的是,在运用上述自适应算法进行模拟时,取 $\hat{\sigma}=0.2$ 的原因是由于初始值 σ_0 得到的样本性质已经比较好,所以我们估计较合适的 σ 值与初始值 σ_0 相差不大.另外,由于 σ 值一定是正数,故在自适应算法的第 6 步中以 σ_n 为中心进行随机游动时,若得到的是一个负数,则必须舍去.

基于以上得到的样本,我们可以对密封圈失效与温度之间的关系作出一些粗略的推断.图 6 为根据样本估计得到的在不同温度下密封圈失效的概率密度.易见,温度为华氏 50 度时密封圈失效的概率(大部分集中在"1"即失效附近)要明显大于温度为华氏 70 度时密封圈失效的概率,从而暗示导致密封圈失效的原因很有可能是寒冷的天气.

万方数据

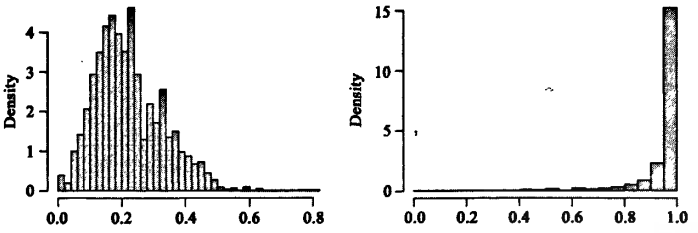

图 6 左图:温度为华氏 70 度时,根据样本估计得到的密封圈失效的概率密度;右图:温度为华氏 50 度时,根据样本估计得到的密封圈失效的概率密度.

5 结论

本文主要对 MCMC 方法中一类很重要的算法、即 Metropolis-Hastings 算法作了分析和改进.从理论上来说,M-H 算法构造的 Markov 链具有可逆性,从而保证了其唯一的极限分布为目标分布.从实际计算结果来看,在提议函数选择合适的情况下,用文中提出的 M-H 自适应算法模拟得到的样本能够很好地反映目标分布的特性.虽然并非在任何情况下算法的效率都很高,但正如文中所述,只要我们能够对算法进行一些实验分析,我们就有可能找到合适的提议函数使得算法的效率大大提高.

MCMC 方法对统计推断的发展的贡献是巨大的,尤其是在贝叶斯模型中,MCMC 方法解决了长久以来一直困扰贝叶斯模型求解的计算问题.本文以贝叶斯 Logistic 模型为例说明了 M-H 自适应算法在贝叶斯模型中的应用.从分析结果上看,以后验分布为依据作出的统计推断与实际情况吻合的较好.MCMC 方法是一种正在不断发展完善的计算方法.本文对 M-H 算法做了比较分析及改进讨论,这将有助于在实践中寻找合适高效的算法用于解决实际问题.

参考文献:

[1] Altaleb A, Chauveau D. Bayesian analysis of the Logit model and comparison of two Metropolis-Hastings strategies[J]. Computational Statistics & Data Analysis, 2002,39(1): 137 - 152.

[2] Roberts G O, Rosenthal J S. Optimal scaling for various Metropolis-Hastings algorithms[J]. Statistical Science, 2001, 6(4):351 - 367.

[3] Geweke J, Tanizaki H. Note on the sampling distribution for the Metropolis-Hastings algorithm[J]. Communication in Statistics-Theory and Methods, 2003, 32(4): 775 - 789.

[4] Sawyer S. The Metropolitan-Hastings algorithm and extensions[J]. Washington University, April 17, 2004.

[5] 孟庆芳,张强,牟文英. 混沌序列自适应多步预测及在股票中的应用[J]. 系统工程理论与实践,2005,25(12): 62 - 68.
Meng Q F, Zhang Q, Mu W Y. A novel adaptive multi-step-prediction met hod for chaotic time series and its applications in stock market[J]. Systems Engineering - Theory & Practice, 2005,25(12): 62 - 68.

[6] Chen P. Some nonparametric estimators and their properties under the competing risks case[J]. Sankhya: Indian J Statist Series A, 1998, 60(2): 293 - 304.

[7] Chen P. Estimators and some behaviors for a partially linear model with censored data[J]. Acta Mathematica Scientia, 1999,19 (3): 321 - 331.

[8] Chen P, Yan F R, Wu Y Y, et al. Detection of outliers in ARMAX time series models[J]. The 5th IIGSS Workshop, Wuhan. June, 2007, to appear.

[9] Chen P, Chen Y. The Identification of Outliers in ARMAX Models via Genetic Algorithm[J]. The 5th IIGSS Workshop, Wuhan June, 2007, to appear.

[10] 陈平,达庆利. 运用 SAS 软件系统对我国农作物受灾及成灾面积的预测分析[J]. 系统工程理论与实践,2001,21(4): 141 - 144.
Chen P, Da Q L. Prediction and analysis for the disaster area of crops in our country by SAS system[J]. Systems Engineering - Theory & Practice,2001,21(4): 141 - 144.

[11] 陈平,王成震,周俊,等. 运用 SAS 软件对上证指数月线数据的综合预测分析[J]. 系统工程理论与实践, 2003, 23 (6): 36 - 41.
Chen P, Wang C Z, Zhou J, et al. Prediction and analysis for the index of Shanghai stock exchange by SAS system on monthly data [J]. Systems Engineering - Theory & Practice, 2003, 23(6): 36 - 41.

万方数据

Metropolis-Hastings自适应算法及其应用

万方数据 WANFANG DATA 文献链接

作者： 陈平， 徐若曦， CHEN Ping， XU Ruo-xi
作者单位： 陈平,CHEN Ping(东南大学,数学系,南京,210096)， 徐若曦,XU Ruo-xi(俄亥俄州立大学,统计系,美国,43210-1247)

刊名： 系统工程理论与实践 ISTIC EI PKU
英文刊名： SYSTEMS ENGINEERING-THEORY & PRACTICE
年，卷(期)： 2008，28(1)
被引用次数： 0次

参考文献(11条)

1. Altaleb A. Chauveau D Bayesian analysis of the Logit model and comparison of two Metropolis-Hastings strategies 2002(01)

2. Roberts G O. Rosenthal J S Optimal scaling for various Metropolis-Hastings algorithms 2001(04)

3. Geweke J. Tanizaki H Note on the sampling distribution for the Metropolis-Hastings algorithm 2003(04)

4. Sawyer S The Metropolitan-Hastings algorithm and extensions 2004

5. 孟庆芳. 张强. 牟文英 混沌序列自适应多步预测及在股票中的应用[期刊论文]-系统工程理论与实践 2005(12)

6. Chen P Some nonparametric estimators and their properties under the competing risks case 1998(02)

7. Chen P Estimators and some behaviors for a partially linear model with censored data[期刊论文]-Acta Mathematica Scientia 1999(03)

8. Chen P. Yan F R. Wu Y Y Detection of outliers in ARMAX time series models 2007

9. Chen P. Chen Y The Identification of Outliers in ARMAX Models via Genetic Algorithm 2007

10. 陈平. 达庆利 运用SAS软件系统对我国农作物受灾及成灾面积的预测分析[期刊论文]-系统工程理论与实践 2001(04)

11. 陈平. 王成震. 周俊 运用SAS软件对上证指数月线数据的综合预测分析[期刊论文]-系统工程理论与实践 2003(06)

相似文献(10条)

1. 期刊论文 朱嵩. 毛根海. 刘国华. 黄跃飞. ZHU Song. MAO Gen-hai. LIU Guo-hua. HUANG Yue-fei 改进的MCMC方法及其应用 -水利学报2009,40(8)

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.

2. 期刊论文 王燕萍. 吕震宙. 赵新攀. WANG Yan-ping. LV Zhen-zhou. ZHAO Xin-pan 基于MCMC的PLP未来强度的Bayesian预测分析 -航空计算技术2010,40(2)

在无信息先验分布下,将Gibbs抽样与Metropolis-Hastings算法相结合的方法应用于幂律过程的未来强度的Bayesian预测.该预测方法能将时间截尾数据和失效截尾数据统一分析,并给出在未来某一时刻处强度函数的MCMC样本,利用该样本可以方便地获得关于未来某一时处刻强度函数及其函数的各种后验分析.一个经典工程数值算例说明了预测方法的可行性、合理性与有效性.

3. 期刊论文 白伟. 何晨. 诸鸿文 MCMC方法及其在移动通信中的应用 -通信技术2002,""(8)

马尔可夫链蒙特卡罗(Markov Chain Monte Carlo,MCMC)方法是从统计科学发展而来的、可以提供理论上的最优性能、计算量较低的一种非常有应用前景的新型信号处理方法.文中结合移动通信中的应用,简单介绍MCMC的基本原理,并总结出MCMC的一些特点.

4. 学位论文 李萌 ARMA-GARCH-M模型的马氏链抽样算法与实证分析 2002

为了将市场风险更好地反应在投资回报中,Engle等人(1987)引入了GARCH-M模型.作为该模型的推广,我们在该文中提出了一个一般的ARMA(p,q)-GARCH(r,s)-M(k)模型,并在详细给出模型的后验分布以及模型的所有参数的满条件分布的基础上,结合Chib and Greenberg(1994)与Nakatsuma(2000)等人的工作,对此新模型设计了一个可行的混合Metropolis-Hastings算法,简化了MA块与GARCH块的估计.同时,使用该算法对上证指数(1998年1月1日至2001年12月31日)的收益率的AR-GARCH-M模型的参数进行了估计,并分析了数据的整合性与市场的风险效应.在论文的第二大部分内容(第五章)中,我们独立地利用经典的统计方法,用两个模型(IGARCH(1,1)-M模型和EGARCH(1,1)-M模型)对中国股票市场的风险特征加以讨论.通过统计分析得到,对上证指数的收益率,IGARCH(1,1)-M模型与EGARCH(1,1)-M模型的统计描述效果基本相同,其结果各有千秋;面对深圳成份指数的收益率,IGARCH(1,1)-M模型的描述结果

要略好于EGARCH(1,1)-M模型.这个结果对研究中国证券市场的风险补偿和波动率依赖的特征具有重要的意义.

5.期刊论文 包云霞.鲁法明.刘福升.BAO Yun-xia.LU Fa-ming.LIU Fu-sheng 非线性状态空间模型的EHMM抽样法 -山东科技大学学报（自然科学版）2005,24(3)

给出了一种新的非线性状态空间模型的MCMC方法--EHMM(Embedded Hidden MarkovModel)抽样法,运用该方法构造的Markov链的收敛速度比传统的MCMC方法有明显提升,文中证明了这一结论并以一维非线性状态空间模型为例加以说明.

6.期刊论文 王亚民.孙长森.汤在祥.胡治球.徐辰武.WANG Ya-min.SUN Chang-sen.TANG Zai-xiang.HU Zhi-qiu.XU Chen-wu 谷物胚乳性状QTL区间作图的贝叶斯方法 -扬州大学学报（农业与生命科学版）2008,29(3)

将贝叶斯统计原理和胚乳性状的数量遗传模型相结合,以分离群体中各植株的分子标记基因型以及植株上若干粒种子胚乳性状的单粒观测值为数据模式,提出胚乳性状QTL区间作图的贝叶斯方法.该方法通过Gibbs以及Metropolis-Hastings抽样实现的马尔科夫链蒙特卡罗(MCMC)算法获得QTL效应和位置的估计.方法的有效性用染色体水平和基因组水平2套模拟方案进行验证,结果表明:贝叶斯方法能够准确地估计胚乳性状QTL的位置和效应,并同时区分2种显性效应.

7.学位论文 蔡俊娟 SLTBL模型的Bayes估计与子集选择 2006

可分离的下三角双线性模型是一类既具有广泛性又具有良好概率结构的双线性模型，简记为”SLTBL模型”.本文采用Bayes方法对可分离的下三角双线性模型进行了统计分析.通过设置合理的先验，得到了各个参数的联合后验密度，进而导出了所有参数的条件后验分布.我们利用Gibbs抽样器方法抽取后验密度的样本，以对参数进行统计推断.特别地，由于从模型的方向向量的条件后验分布中直接抽样是困难的，我们特别设计了一个简单有效的Metropolis-Hastings算法以解决该难题.进一步，我们采用逆跳MCMC技术，以解决模型的定阶及子集选择问题.我们用仿真例子演示了本文所建议的方法，并应用于分析实际数据.

8.期刊论文 薛亚茹.李明 一种有效的粒子滤波器的改进算法 -电子元器件应用2008,10(5)

为了解决粒子滤波算法中重采样后粒子中包含重复点过多,从而丧失了粒子的多样性等问题,文中在重采样后引入一个马尔可夫链蒙特卡罗(MCMC)移动步骤来增加粒子的多样性,因而能更好地近似状态的后验概率分布.

9.学位论文 朱嵩 基于贝叶斯推理的环境水力学反问题研究 2008

随着环境水力学正问题模型(预测模型)求解的日益成熟，其有关的反演题的研究越来越受到了的重视。环境水力学反问题解的主要困难在于其不适定性，而这种不适定性(尤其是不唯一性)主要来源于水环境系统的不确定性。水环境系统中含有许多不确定性因素，如测量数据的不确定性、数学模型的不确定性和模型参数的不确定性等，如何正视并解决这些不确定性是环境水力学反问题研究亟需解决的问题。本文从概率论的角度出发采用贝叶斯统计学的方法-贝叶斯推理建立了环境水力学反问题的求解模型，其中模型参数、测量数据、先验信息和最终反问题解的形式等全部采用概率语言描述，这样就较好她解决了环境水力学反问题求解中的不确定性问题。

理论研究方面，本文对贝叶斯推理中标准的抽样技术-马尔科夫链蒙特卡罗抽样方法(Markov chain Monte Carlo，MCMC)进行了研究。为了提高马尔科夫链对后验参数空间(解空间)的搜索能力，本文在Metropolis-Hastings算法的核心上提出了一种动态多链的搜索方法Dynamic Multi-chains Metropolis-Hastings，DMMH)。该算法由于采用多条链的搜索，因而增加了对后验空间的搜索能力，克服了单链搜索能力不足的问题；同时DMMH中马尔科夫链可以根据实时的搜索状态进行调整(减少)链的个数，因而比常规的多链搜索节省了计算时间。

环境水力学正问题求解的基础数学模型为对流扩散方程，本文主要采用有限体积法和有限元法作为求解方法。在有限体积法方面，特别研究了基于通量限制器(flux limiter)的有限体积法，该方法具有精度商、稳定性好的优点，比较适用于需要调用大量正演计算的环境水力学反问题的求解。本文提出了一种改进的通量限制器格式M7，数值计算表明M7通量限制器与传统的一些限制器相比，提高了计算精度和显式差分计算的稳定性。

应用研究方面，本文采用贝叶斯推理主要求解了两类反问题，即环境水力学的参数估计反问题和污染物源项识别反问题。环境水力学数学模型丰富，为了验证基于贝叶斯推理的反演模型的合理性和可靠性，本文根据模型分类的不同，研究了一维模型、二维模型、稳态模型、非稳态模型、常系数模型、变系数模型、含源模型、非含源模型等各种类型的环境水力学数学模型，大量算例计算表明在正问题求解精度高(数值误差低)的情况下，采用贝叶斯推理和MCMC抽样方法获得的反问题的解具有信息量大(能给出环境水力学参数的后验分布)且估计精度高的优点。在反演结果的可靠性和估计的精度方面，本文重点探讨了测点位置、测量异常值和测点数目对反演结果的影响，研究了反问题求解的非唯一性问题。源项识别反问题主要研究了单个污染源和多个污染源下的位置和强度的识别问题，采用MCMC的贝叶斯推理均能获得较高的识别精度。

最优化方法目前是环境水力学反问题求解的主流方法，它和贝叶斯方法两者之间既有区别也有联系。本文也介绍了基于优化方法(遗传算法)的环境水力学反问题求解方法，并提出了基于混合遗传算法结合有限体积法FVM-HGA)的参数识别方法。该方法与传统遗传算法相比具有较强的全局寻优和局部搜索能力，且反演的精度高、收敛速度快；但其缺点是不能像贝叶斯方法那样获得参数的后验分布。

由于MCMC本身也可作为参数识别的工具，本文也对基于MCMC. 的环境水力学参数识别问题进行了研究，提出了基于贝叶斯理论的以似然函数为收敛准则的参数识别新方法(L-MCMC)。算例表明，L-MCMC在减少反演计算量上效果显著。

总而言之，对于诸如环境水力学反问题等一些环境系统反问题，基于MCMC的贝叶斯方法是一种强大的求解工具，既能给出模型参数的分布规律，又能作为优化算法给出模型参数的最优估计，具有重要的研究价值和应用前景。

10.会议论文 汤在祥.徐辰武 基于贝叶斯统计的遗传连锁分析方法 2005

统计分析在遗传学研究领域有着广泛的应用.然而,大多数的研究人员都使用经典的统计分析方法,包括:假设检测、参数的点估计和区间估计等.这些经典方法在多数情况是十分有效的.然而,近年来贝叶斯统计以其鲜明的特点和独到的分析方法,已逐渐引起了人们的重视,在统计学界逐渐被认可.Gelman等、David、Chib等以及Hastings详细阐述了马尔可夫链蒙特卡罗理论(MCMC)、Gibbs抽样以及Metropolis-Hastings算法,使得贝叶斯方法能够很好实现.对侧重于统计应用的研究者来说,它能为一些问题提供更直接的解决方法并可将先验信息综合其中;同时,贝叶斯方法对结果的解释更加直观.特别是在复杂的遗传学问题或者经典的统计方法无法解决的问题上,贝叶斯统计已得到迅速发展,并日益得到遗传学家的广泛应用.遗传图谱在遗传学的诸多领域,如QTL分析,图位克隆方面起着重要作用.而图谱的构建有赖于基因座位间的遗传重组率或遗传距离的估计,这是遗传学研究中的经典问题,也是基因组分析的基础.本文从简单的两点和三点分析出发,探讨了贝叶斯理论和MCMC模型在连锁分析和遗传图谱构建方面的应用.在本文两个位点的连锁分析中,如何选取合适的先验分布是一个重要问题,因为先验的选取对后验分布影响很大.而选取p分布作为遗传重组率分布的先验分布,这是因为两个位点的遗传重组率的参数r在以往的诸多研究中很容易获得,如果现在获得了一个观察数据,要得到基于这个观察数据的重组率,它显然有一个客观、合理的先验分布,就是一个β分布.选取这样的分布符合贝叶斯先验分布选取的基本原则,同时Gelman等人也建议使用β分布作为遗传重组率分布的先验分布.导出r的先验分布后就可获得其后验分布,采用Gibbs抽样方法在其后验分布中抽取大量样本,并进行统计分析,这样即可得到遗传重组率的后验平均值和标准差.然而,构建连锁图谱首先需要测验两个标记或基因位点是否在同一连锁群上.因此,对于上面获得的遗传重组率需要进行假设检验.这里可以求得相应贝叶斯因子,作为两个标记或基因位点连锁与否的判别标准.由此,推广到多个位点情况,就可以确定一个连锁群.确定连锁群后,需要对连锁群上所有位点进行排序.本文以简单的三个位点情况进行分析.以β分布作为遗传重组率分布的先验分布,在约束符合系数在0到1的情况下,采用基于Metropolis-Hastings算法的MCMC模型,通过大量的运算,获得连锁群上各个基因位点最可能的排序,并计算各位点间的遗传距离,从而构建相应的遗传图谱.本文虽然只是研究了两点和三点的情况,但这完全可以推广到一个连锁群的情况,实现基于贝叶斯统计的遗传连锁图谱的构建.为说明该方法的有效性和实用性,我们编制了相应的SAS程序,将模拟研究和实例分析的结果与极大似然法进行了比较,结果表明贝叶斯方法显然是有效而实用的.

本文链接：http://d.wanfangdata.com.cn/Periodical_xtgcllysj200801015.aspx
授权使用：辽宁师范大学(lnsfdx)，授权号：21b35f64-367f-49f5-a94a-9ddf008ea584

下载时间：2010年8月27日