Mathematical Modeling
(for psychology students…)

Approximate inference (2)—MCMC2

张洳源

2024/04/11

上海交通大学心理与行为科学研究院
上海交通大学医学院附属精神卫生中心

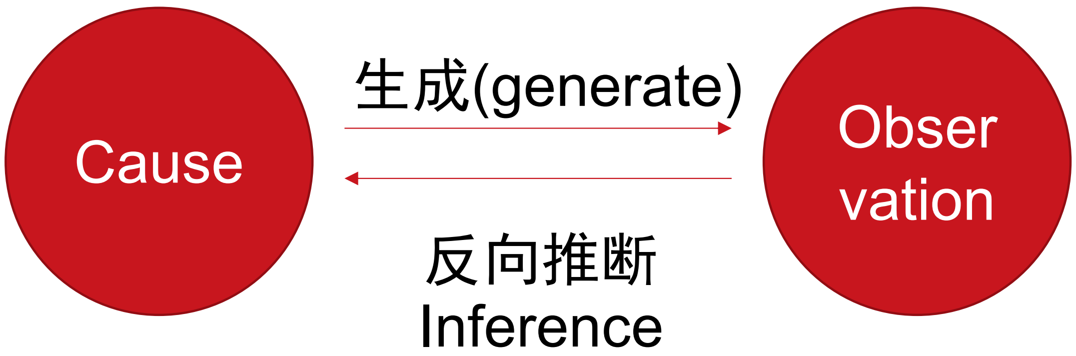
Outlines

- What have we learned
  - Markov Chain，Monte Carlo Methods，Sampling
  - Metropolis-Hasting Algorithm的推导
  - 手写 Metropolis-Hasting Algorithm

- What will we learn today?
  - MCMC for parameter estimation
    - Estimate psychometric function
  - MCMC for theories in Psychology
    - The Bayesian Brain
    - Population codes

Generative Process and Reverse Inference

- Generative process (由原因生成观察数据)
- Reverse Inference (由观察数据反推原因)



频率学派 vs. 贝叶斯学派

频率学派 ← 最大似然估计

参数点估计

Exact Inference

最大后验估计

贝叶斯学派

马尔科夫链蒙特卡洛采样 (MCMC)

变分推断 (Variational Inference)

贝叶斯参数估计

Approximate Inference

马尔科夫链蒙特卡洛采样 (MCMC)和变分推断(Variational Inference)后面的内容会说到

Bayesian Theory

$$\overset{\text{后验分布}}{p(\theta|d)} = \frac{\overset{\text{似然函数}}{p(d|\theta)} * \overset{\text{先验分布}}{p(\theta)}}{p(d) \longrightarrow \text{归一化因子}}$$

有时候写成

$$p(\theta|d) \propto p(d|\theta) * p(\theta)$$

认知心理学中概率模型的本质就是求解后验概率分布$p(\theta|d)$

MCMC sampling in psychology

- MCMC sampling as a technique for parameter estimation

- MCMC sampling as a cognitive theory

Recap

总体目标，求后验概率$p(\theta|data)$

怎么求？从后验概率$p(\theta|data)$分布采很多样本去近似

Markov chain

任何一个分布，都有其对应的一个平稳细致转移函数(或矩阵)

从分布的任意状态出发，沿着转移函数(或矩阵)进行转移，就可以对分布采样

任意一个转移函数Q，满足细致平稳条件的转移矩阵

$$M = \mathrm{Q} * \alpha = Q * \min\left[1, \frac{p(\mathrm{d} \mid x^*)}{p(d \mid x_t))}\right]$$

Monte Carlo

避免求积分的简单粗暴的近似思想

Sampling

逆变换采样
(常见分布)

接受拒绝采样
(复杂分布，但需要常数C)

先采样再按比例接收的思想

Metropolis-Hasting算法

1. 从任意一个x_0开始
2. 根据当前的样本x_t构建一个提议分布$g(x|x_t)$，然后采样得到x^*
3. 然后计算接受率 $\alpha = \min\left[1, \frac{p(\mathrm{d} \mid x^*)}{p(d \mid x_t))}\right]$
4. 从[0,1]采样一个值u,
   - 如果 u< α, 接受该样本$x_{t+1} = x^*$
   - 否则，继续维持原来的样本$x_{t+1} = x_t$
5. 重复2-4步直到结束

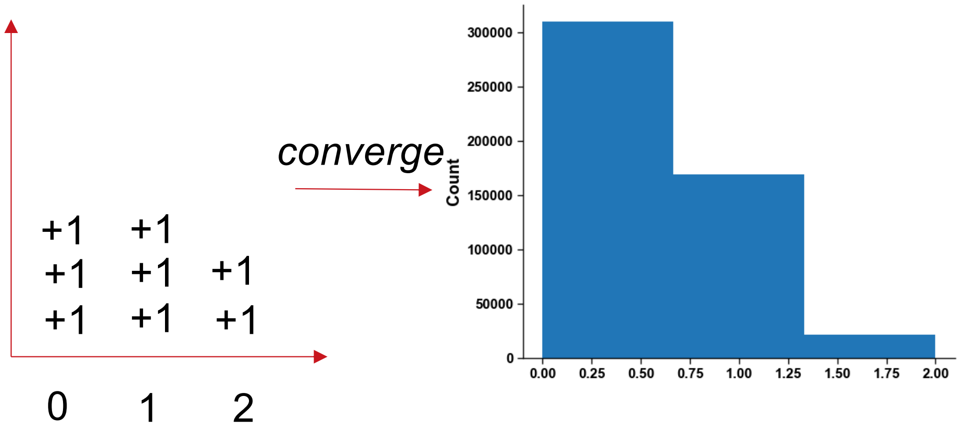
MCMC

三个离散状态: 0, 1, 2

马尔科夫链对应的转换矩阵*P*

随便指定一个初始状态

$$0 \xrightarrow{P} 1 \xrightarrow{P} 2 \xrightarrow{P} 1 \xrightarrow{P} 2 \xrightarrow{P} 0 \xrightarrow{P} 0 \xrightarrow{P} 1$$

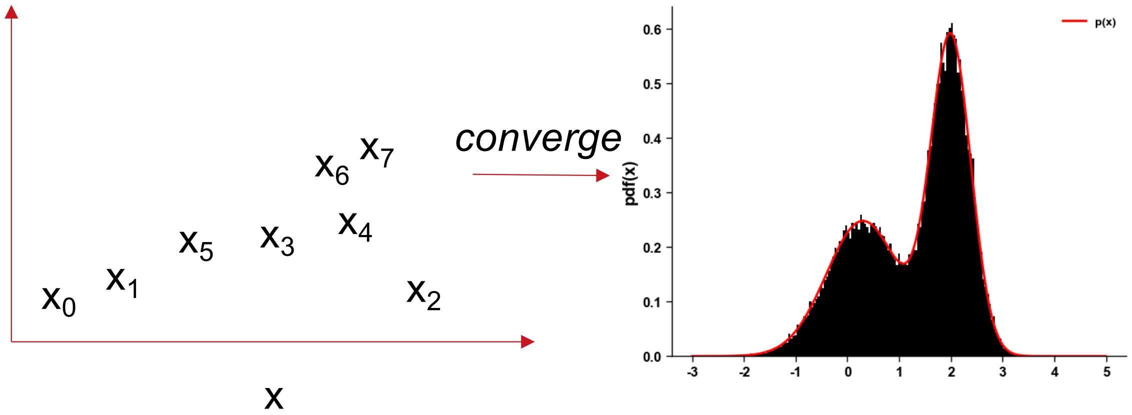


无限个连续状态 $x_0, x_1, x_2 \ldots$

马尔科夫链对应的转换函数Q

随便指定一个初始状态

$$x_0 \xrightarrow{Q} x_1 \xrightarrow{Q} x_2 \xrightarrow{Q} x_3 \xrightarrow{Q} x_4 \xrightarrow{Q} x_5 \xrightarrow{Q} x_6 \xrightarrow{Q} x_7$$



Metropolis-Hasting sampling

对于任意一个分布$p(x)$，以及任意一个转移函数$Q(x_j|x_i)$，我们可以构建一个一定满足细致平稳条件的新转移函数

$$M(x_j|x_i) = Q(x_j|x_i) * min(1, \frac{p(x_j) * Q(x_i|x_j)}{p(x_i) * Q(x_j|x_i)})$$

当 $p(x_j) * Q(x_i|x_j) \geq p(x_i) * Q(x_j|x_i)$ 有 $M(x_j|x_i) = Q(x_j|x_i)$

直接用任意构建的$Q(x_j|x_i)$ 来进行下一个采样

当 $p(x_j) * Q(x_i|x_j) < p(x_i) * Q(x_j|x_i)$ 有 $M(x_j|x_i) = Q(x_j|x_i) * \frac{p(x_j) * Q(x_i|x_j)}{p(x_i) * Q(x_j|x_i)}$

先用任意构建的$Q(x_j|x_i)$ 来进行下一个采样，然后再以$\frac{p(x_j)*Q(x_i|x_j)}{p(x_i)*Q(x_j|x_i)}$的概率来决定是否接受

Metropolis-Hasting sampling

0, 从任意一个x_0开始

1, Q为高斯函数(均值为x_0，std自定)，从中采样得到x_1，

2, 用$\alpha \sim (0,1) < \min(1, \frac{p(x_1)}{p(x_0)})$来判断是否接受$x_1$，是，接受$x_1$；

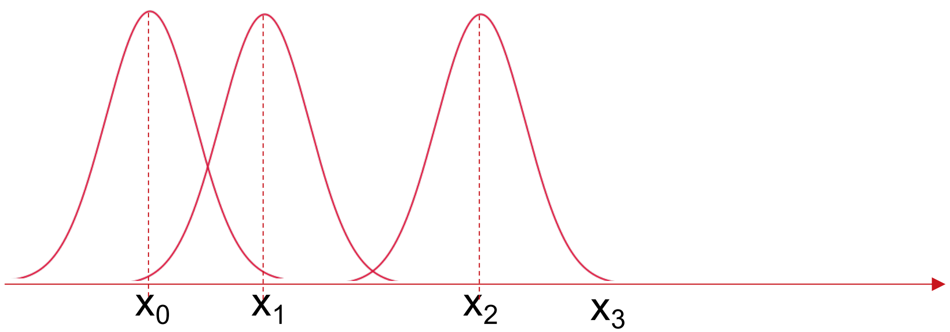
3. Q为高斯函数(均值为x_1，std同上)，从中采样得到x_2，

4. 用$\alpha \sim (0,1) < \min(1, \frac{p(x_2)}{p(x_1)})$来判断是否接受$x_2$，是，接受$x_2$；

5. Q为高斯函数(均值为x_2，std同上)，从中采样得到x_3，

6. 用$\alpha \sim (0,1) < \min(1, \frac{p(x_3)}{p(x_2)})$来判断是否接受$x_3$，是，接受$x_3$；

。。。



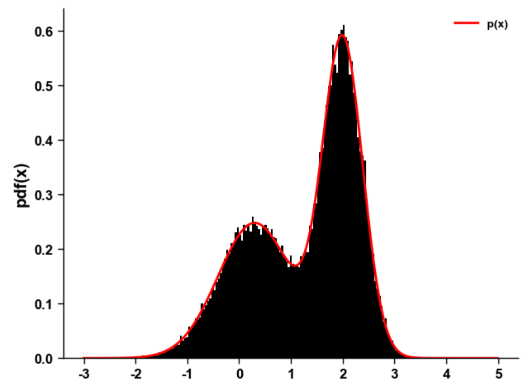
无限个连续状态 $x_0, x_1, x_2 \ldots$

马尔科夫链对应的转换函数M

随便指定一个初始状态

$$x_0 \xrightarrow{M} x_1 \xrightarrow{M} x_2 \xrightarrow{M} x_3 \xrightarrow{M} x_4 \xrightarrow{M} x_5 \xrightarrow{M} x_6 \xrightarrow{M} x_7$$

$\xrightarrow{\textit{converge}}$



MCMC for parameter estimation

RESEARCH

RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Computational and neurobiological foundations of leadership decisions

Micah G. Edelson*, Rafael Polania, Christian C. Ruff, Ernst Fehr*, Todd A. Hare*

For latent variables at the highest level of the hierarchy (hyper-group parameters), we assumed flat uninformed priors (i.e., uniform distributions). Posterior inference of the parameters in the hierarchical Bayesian models was performed via the Gibbs sampler using the Markov Chain Monte Carlo (MCMC) technique implemented in JAGS (*70*, *71*). A total of 50,000 samples were drawn from an initial *burn-in* sequence, and subsequently a total of 50,000 new samples were drawn using three chains (each chain was derived based on a different random number generator engine, using a different seed). We applied a *thinning* of 50 to this sample, resulting in a final set of 1,000 samples for each parameter. This thinning assured that the final samples were not auto-correlated for all of the latent variables of interest investigated in the study. We conducted Gelman-Rubin tests for each parameter to confirm convergence of the chains. All latent variables in our Bayesian

编程练习

- MCMC方法在认知计算中的应用
  - 用MLE估计心理物理曲线的参数
    - 缺点: 使用MLE的方法，中间的隐变量最好是能解析解积分消除，否则求解很麻烦

见7mlepsymetric.ipynb

编程练习

- MCMC方法在认知计算中的应用
  - 用MLE估计心理物理曲线的参数
    - 缺点: 使用MLE的方法，中间的隐变量最好是能解析解积分消除，否则求解很麻烦
  - 手写MH算法来求解心理物理曲线的参数

编程练习

- MCMC方法在认知计算中的应用
  - 用MLE估计心理物理曲线的参数
    - 缺点: 使用MLE的方法，中间的隐变量最好是能解析解积分消除，否则求解很麻烦

  - 手写MH算法来求解心理物理曲线的参数

见8mcmcpsymetric.ipynb

注意点:

1. 如果采样的分布限制在一定范围，比如(0,1)，请先用高斯转移函数在实数范围内采样，然后把样本通过logistic函数转成(0,1)的范围

2. 直接求$\min(1, \frac{p(x_i)}{p(x_j)})$不太好求，可以转换成$\min(0, \log\big(p(x_i)\big) - \log(p(x_j)))$

编程练习

- MCMC方法在认知计算中的应用
  - 用MLE估计心理物理曲线的参数
    - 缺点: 使用MLE的方法，中间的隐变量最好是能解析解积分消除，否则求解很麻烦
  - 手写MH算法来求解心理物理曲线的参数
  - 用pymc来求解心理物理曲线的参数

见9pymcpsymetric.ipynb

实现MCMC的工具包

- MCMC的算法本身和具体的模型无关，只是一种采样的方法，所以现在有许多工具包可以实现MCMC的算法。目前的主流MCMC算法包括:
  - Metropolis-Hasting sampler
  - Gibbs sampler
  - Hamiltonian sampler
  - No-U-turn sampler

- 目前的主流Bayesian编程工具包包括
  - Stan (基于C++, 有pystan, rstan接口，推荐)
  - PYMC5 (基于python，不方便写for循环)
  - Numpyro (基于python，不方便写for循环)
  - Jags, winbugs (已被淘汰，不推荐)

一些术语

- Number of MCMC chain
- Burn-in
- Thinning

Background

- 在认知神经科学中，最常用的是MLE, MAP和MCMC

- 相较MLE/MAP, MCMC的优势
  - MLE基于似然函数，很多时候似然函数里面包含多个隐变量和多重积分，无法得到解析形式，无法优化求解。MCMC不需要解析求解任何积分。
  - MLE只是关于参数的点估计。MCMC得到整个后验概率分布，可以得到参数估计的uncertainty
  - MLE很容易会收敛到local minima。MCMC一般不会。

- 相较MLE/MAP, MCMC的劣势
  - 收敛速度慢
  - 得到后验概率之后，不方便做统计以及其他分析。(如何根据sampling近似的后验概率做各种分析是一个热门话题！)

Outlines

- What have we learned
  - Markov Chain，Monte Carlo Methods，Sampling
  - Metropolis-Hasting Algorithm的推导
  - 手写 Metropolis-Hasting Algorithm

- What will we learn today?
  - MCMC for parameter estimation
    - Estimate psychometric function
  - MCMC for theories in Psychology
    - The Bayesian Brain hypothesis
    - Population codes

The "Bayesian brain" Hypothesis

The Bayesian brain considers the brain as a statistical organ of hierarchical inference that predicts current and future events on the basis of past experience.

—Karl Friston

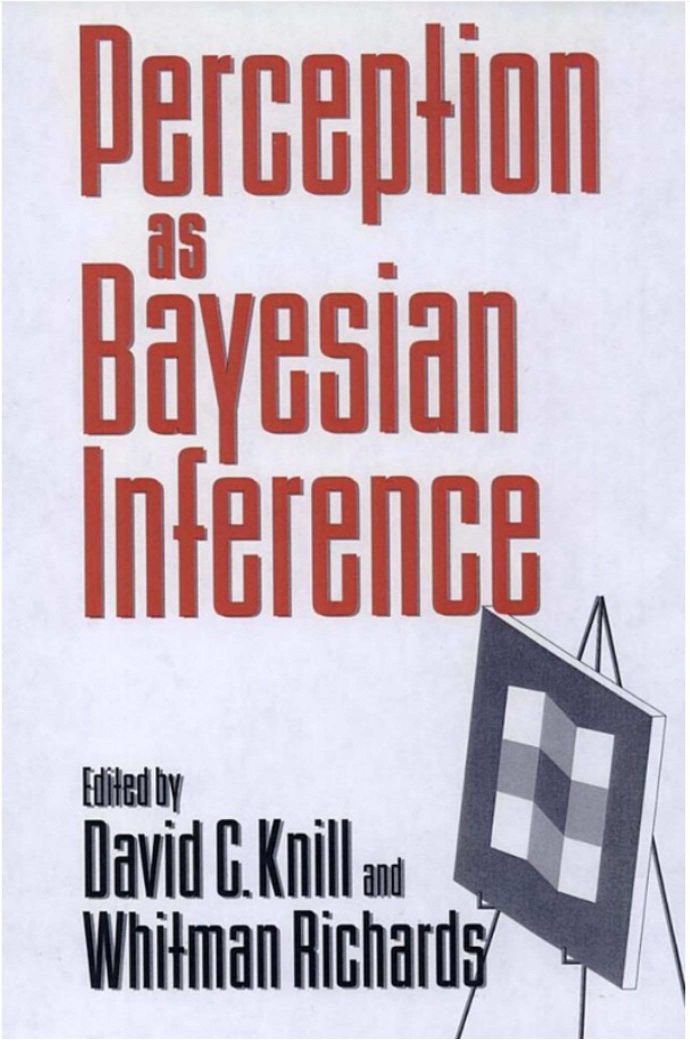
$$p(\theta|d) = \frac{p(d|\theta) * p(\theta)}{p(d)}$$

通俗点讲，Bayesian Brain假说认为大脑是通过贝叶斯推断来完成各种任务

Light-from-above prior

A B

The “Bayesian brain” Hypothesis



The Bayesian revolution in human cognition

Humans are close to ideal Bayesian decision makers

$$p(\theta|d) = \frac{p(d|\theta) * p(\theta)}{p(d)}$$

Knill & Richards, 1996

The “Bayesian brain” Hypothesis

- 如果人脑是按照贝叶斯推断的方式来完成各项认知功能，那么我们脑子里需要表征prior distribution，likelihood function和posterior distribution

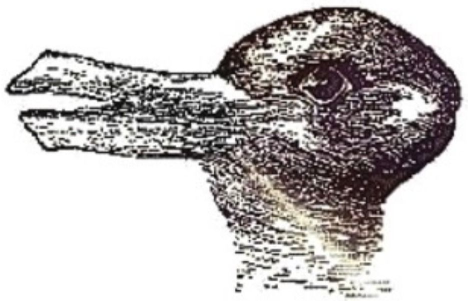
- 人脑如何表征这些成分是当前计算神经科学最前沿的研究之一

$$p(\theta|d) = \frac{p(d|\theta) * p(\theta)}{p(d)}$$

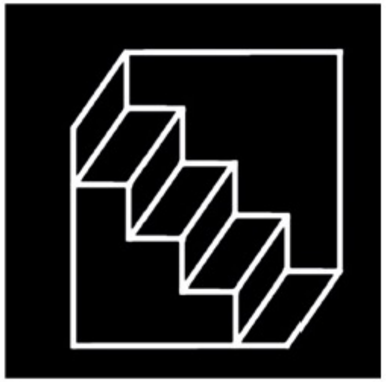
Sampling based inference

Examples of Bistable Figures

A

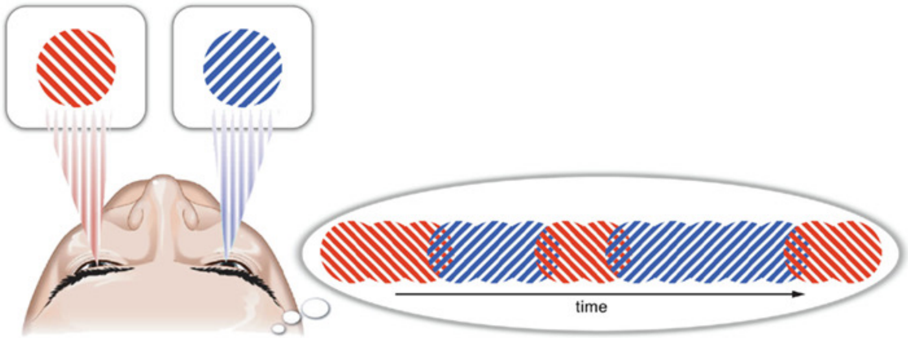


B



C

Binocular rivalry



The “Bayesian brain” Hypothesis

- 人脑如何表征一个概率分布，无论是prior 还是posterior?

- 两种假说：
  - Probabilistic population codes
  - Neural Sampling

The “Bayesian brain” Hypothesis

- 人脑如何表征一个概率分布，无论是prior 还是posterior?

- 两种假说：
  - Probabilistic population codes
  - Neural Sampling

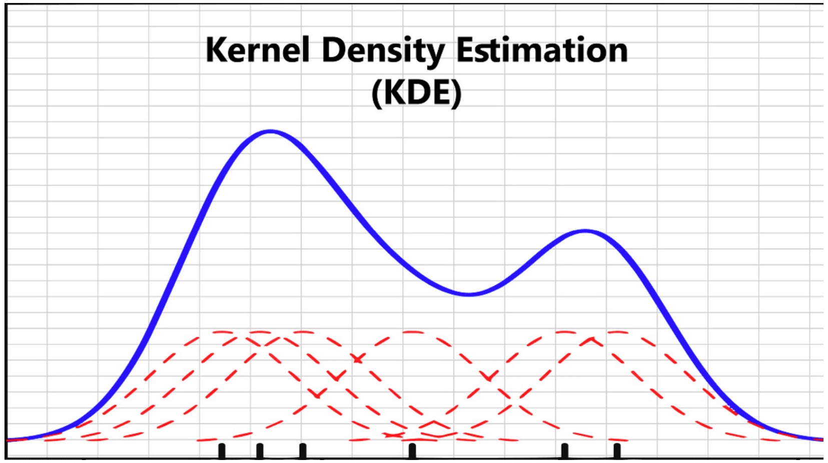
Probabilistic population codes

ARTICLES

nature neuroscience

Bayesian inference with probabilistic population codes

Wei Ji Ma[1,3], Jeffrey M Beck[1,3], Peter E Latham[2] & Alexandre Pouget[1]



每个神经元组成一个kernel(i.e., 基函数)，一群神经元连起来可以表征一个概率分布

Neural sampling

Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior

Patrik O. Hoyer* and Aapo Hyvärinen

Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
http://www.cis.hut.fi/phoyer/
patrik.hoyer@hut.fi

Basis function的坏处：需要的basis function的数量随着分布的维度升高而指数增长

用sampling来近似一个分布，其估计误差只和sample的数量有关，和分布的维度无关！！非常的经济

证明见 https://ruyuanzhang.github.io/resources/Monte%20Carlo%20vs.%20Riemann.html

The “Bayesian brain” Hypothesis

- 人脑如何表征一个概率分布，无论是prior 还是posterior?

- 两种假说：
  - Probabilistic population codes
  - Neural Sampling

这两种假说到底哪个对？如何利用实验证明或者证伪这两种假说是当前计算神经科学的前沿问题之一！！！