自适应抽样算法及其R包开发

ADAPTIVE SAMPLING ALGORITHMS AND ITS R PACKAGE DEVELOPMENT

华东师范大学统计学院　　张东

2017年12月3日

R会议上海场

R User !

2012 旁听R语言会议

2013 第一次学习 R in Action

2014 时间序列分析

贝叶斯统计分析

2015 独立参与商业项目

学校学生讲师

2016 R(武汉场)基础培训

R完成研究生毕业论文

2017 I'm Here~

(伪)随机数生成

常规分布

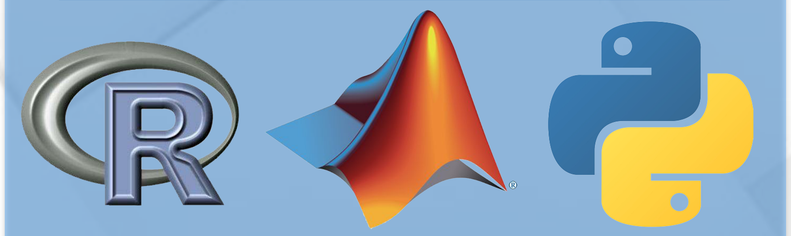
N U Ga

非常规分布

$$f(x) \propto xe^{-(4-x^2)^2}$$

解 决 方 案

成熟的计算机操作技术



算法的理论支持

拒绝抽样法，Mini-Max抽样法，Matropolis-Hasting抽样法，逆变换法，合成法……

背景

速度
+
精度

Adaptive
Algorithm

统计算法

+

软件操作

MARS

ARS

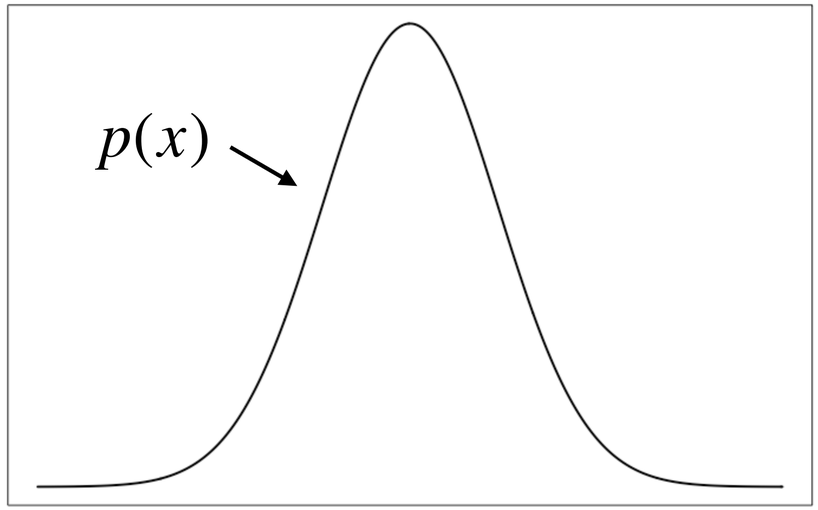
CCARS

ASS

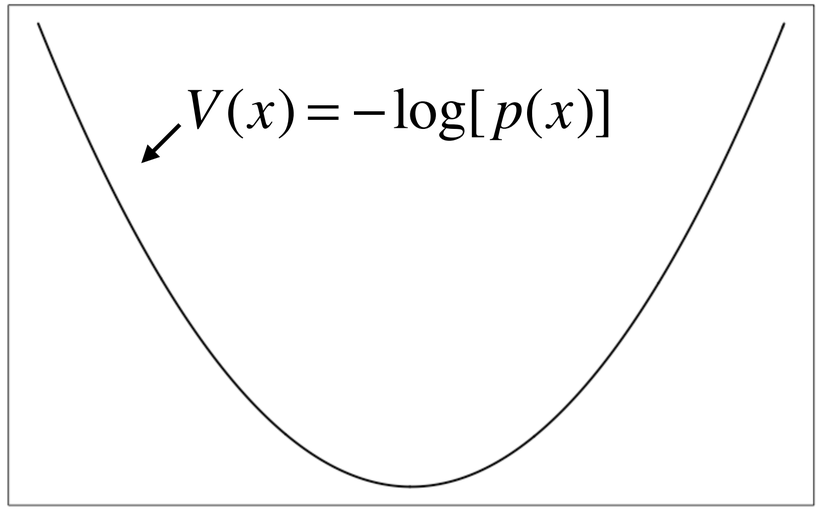
自适应拒绝抽样算法(ARS)

目标概率密度函数$p_0(x) \propto p(x)$，并且记$V(x) = -\log[p(x)]$。

要求：$V''(x) \geq 0$，即$V(x)$为凸(Convex)函数。



(a)



(b)

图1 ARS算法下目标概率密度函数及其负对数变换

人为设定支撑集$S_n = \{s_1, s_2, \ldots, s_{m_n}\}$，在点$(s_k, V(s_k))$做$V(x)$的切线，记作$w_k(x)$。

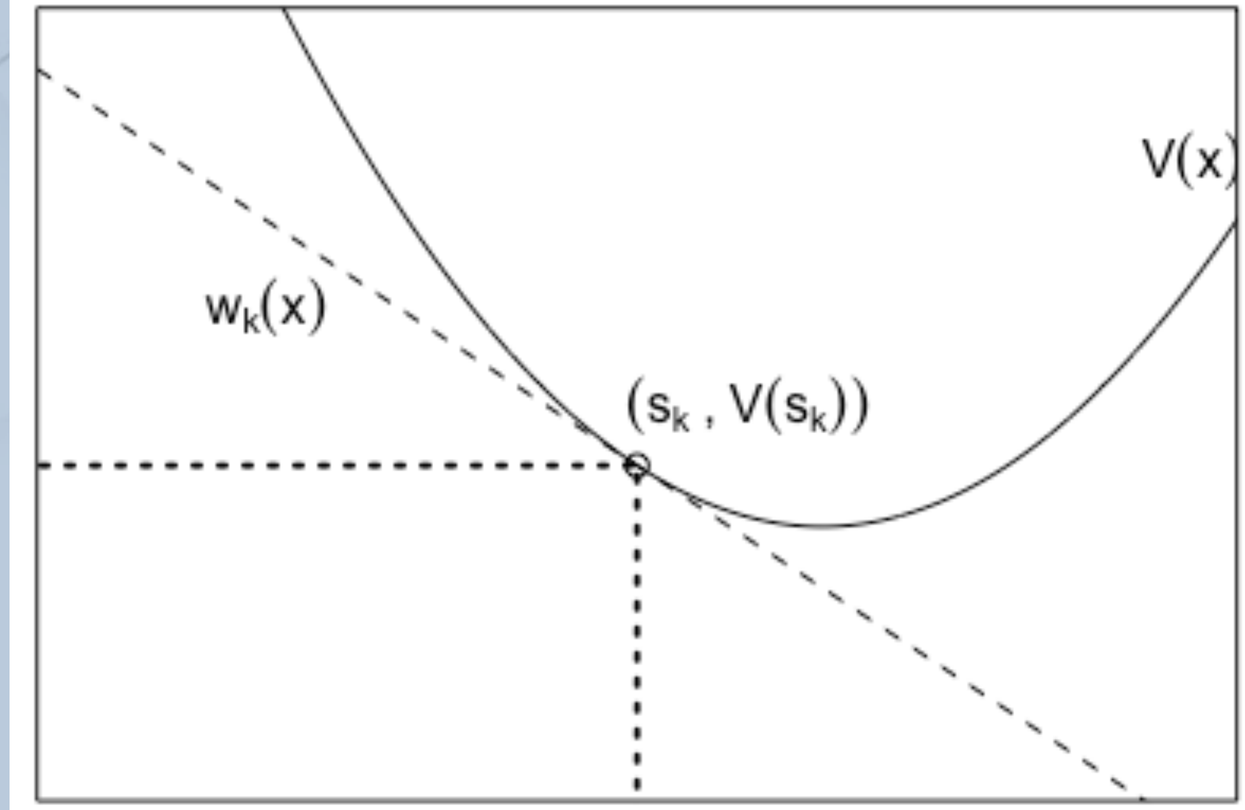


图2 在支撑点处做*V(x)*的切线

令$W_n(x)=\max\{w_1(x),w_2(x),...,w_{m_n}(x)\}$，由构造可知$W_n(x)\le V(x)$。

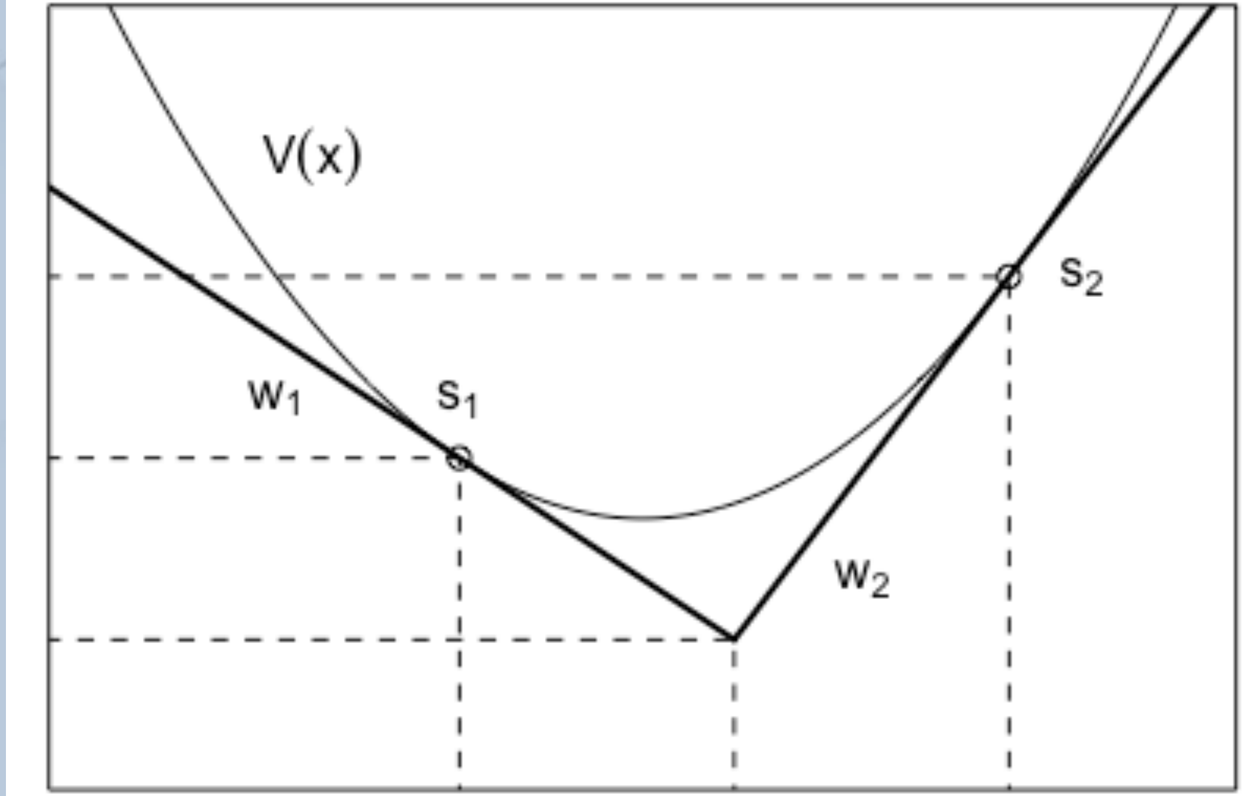


图3 双支撑点时*V(x)*的包络函数$W_2(x)$

记$\pi_n(x) \propto \exp(-W_n(x))$，那么$\pi_n(x)$便是分段指数型分布的概率密度核函数。

Algorithm 1

拒绝步骤

抽取$u \sim U(0,1)$，且从$\pi_n(x)$中生成随机数x'。

如果 $u < \dfrac{p(x')}{\exp(-W_n(x'))}$,则接受x'是从目标分布中生成的随机数。

自适应步骤

如果上拒绝步骤中，x' 被拒绝，那么将其纳入到支撑集中再次抽样。

改进自适应拒绝抽样算法(MARS)

取消了目标概率密度函数为对数凹函数的限制。

操作方法：对$V(x)$的凹凸区间进行划分。

$V(x)$

凸区间

凹区间

» 仿照ARS方法，通过切线方程建立包络函数。

» 使用同一凹区间内相邻支撑点间的割线段构造$V(x)$的包络函数。

如$[x_k, x_{k+1})$为$V(x)$的某个凹区间，$\{z_{k_m}, z_{k_n}\}$为该区间上两个相邻的支撑点，那么从z_{k_m}到z_{k_n}所构造的割线方程(即包络函数)为

$$w_{z_{k_m}, z_{k_n}} = \frac{V(z_{k_m}) - V(z_{k_n})}{z_{k_m} - z_{k_n}}(x - z_{k_m}) + V(z_{k_m})$$

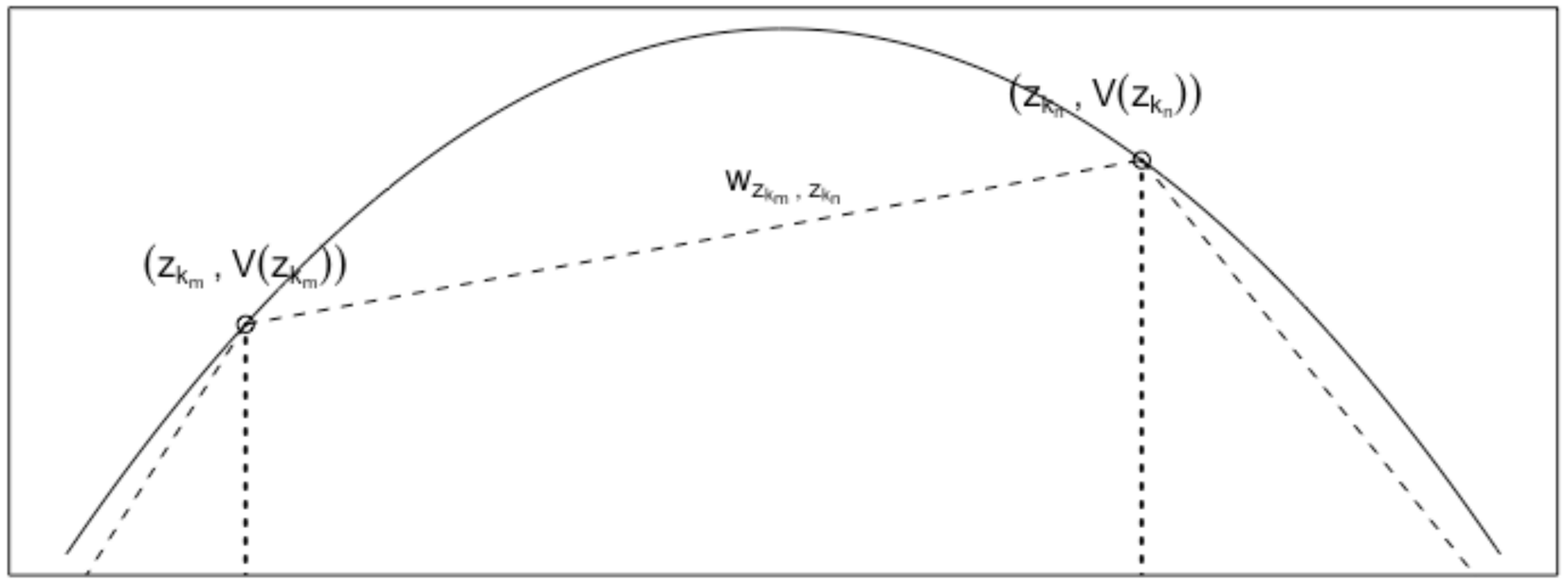


图4　MARS算法下凹区间相邻支撑点构造割线

Algorithm 2

划分步骤　对$V(x)$进行凹凸区间进行划分。
凸区间仿照ARS方法，使用切线构造区间内包络函数；
凹区间使用割线构造区间内包络函数。

凹凸分解自适应拒绝抽样算法(CCARS)

Jordan分解定理的应用

若$f(x)$在区间$[a,b]$上可导且其导函数$f'(x)$是有界变差的，那么其导数$f'(x)$可被分解成一个单调递增函数与一个单调递减函数的和。

推论

函数$f(x)$在区间$[a,b]$上可导且其导函数$f'(x)$是有界变差的，那么$f(x)$可被写成一个凹函数与一个凸函数之和。

Algorithm 3

拆分步骤

对$V(x)$进行拆分$V(x)=V_{\cup}(x)+V_{\cap}(x)$。

对于$V_{\cup}(x)$使用支撑点处切线构造包络函数，记作$W_{\cup}(x)$。

对于$V_{\cap}(x)$使用支撑点处割线构造包络函数，记作$W_{\cap}(x)$。

包络求和步骤

$V(x)$的包络函数为$W(x)=W_{\cup}(x)+W_{\cap}(x)$。

张东

自适应切片抽样算法(ASS)

切片抽样算法 Slice Sampling

通过添加辅助变量达到参数空间扩大的目标，并保持着边际分布与目标分布相同的性质。

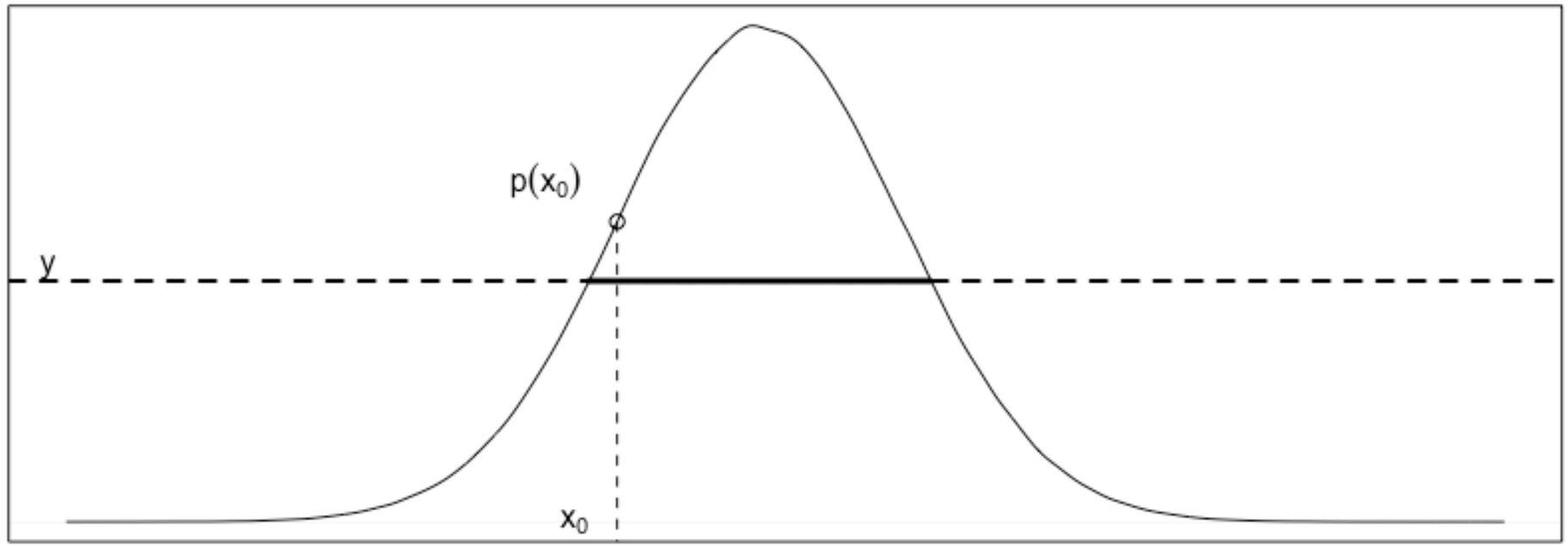


图5 切片抽样算法示意图

自适应切片抽样算法(ASS)的提出
是为了解决切片抽样算法针对多峰分布无效性的问题。

Slice Sampling + Stepping Out >

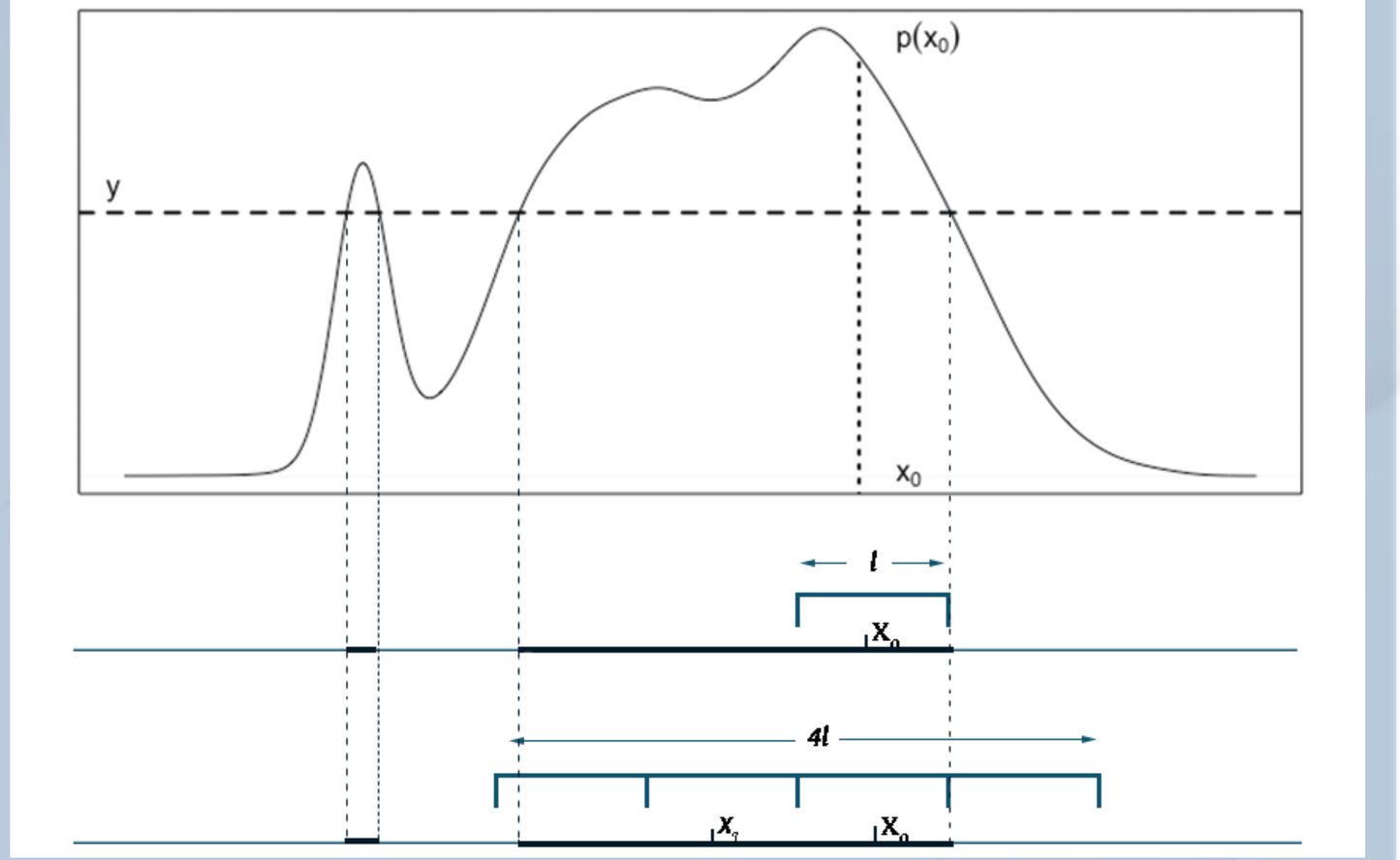


图6 自适应切片抽样算法示意图

> ADAPTIVE ALGORITHM!

张东

Algorithm 4

切片步骤

抽取 $y \sim U(0, p(x_0))$为切片。

覆盖步骤

将覆盖区间w随机放置在x_0附近；

双向等长延拓覆盖区间，直至切片落入区间之内。

自适应步骤

覆盖区间为均匀分布区域抽样z,

如果z落在切片外，那么以z为新端点重新建立抽样区域。

AdapSamp
来啦

https://CRAN.R-project.org/package=AdapSamp

AdapSamp: Adaptive Sampling Algorithms

For distributions whose probability density functions are log-concave, Adaptive Rejection Sampling (ARS) algorithm can be used to build for sampling by Gilks W R, Wild P (1992) <doi:10.2307/2347565>. For others, we can use Modifed Adaptive Rejection Sampling (MARS) algorithm by Martino L, Míguez J (2011) <doi:10.1007/s11222-010-9197-9>, Concave-Convex Adaptive Rejection (CCARS) Sampling algorithm by Görür D, Teh Y W (2011) <doi:10.1198/jcgs.2011.09058> and Adaptive Slice Sampling (ASS) algorithm by Radford M. Neal (2003) <doi:10.1214/aos/1056562461>. So we designed an R package including mainly 4 functions: rARS(), rMARS(), rCCARS(), and rASS(). These functions can realize sampling based on the algorithms above.

Version:	1.0
Imports:	pracma, stats, utils
Published:	2017-09-08
Author:	Dong Zhang
Maintainer:	Dong Zhang <dzhang0716 at 126.com>
License:	GPL-2
NeedsCompilation:	no
CRAN checks:	AdapSamp results

Downloads:

Reference manual:	AdapSamp.pdf
Package source:	AdapSamp_1.0.tar.gz
Windows binaries:	r-devel: AdapSamp_1.0.zip, r-release: AdapSamp_1.0.zip, r-oldrel: AdapSamp_1.0.zip
OS X El Capitan binaries:	r-release: AdapSamp_1.0.tgz
OS X Mavericks binaries:	r-oldrel: AdapSamp_1.0.tgz

Adaptive Sampling

开发者姓名	ZhangDong
开发版本	0.9.0
R的最低版本	3.3.0
使用协议	MIT
依托功能包	HI 与 pracma

函数	功能	参数
rARS( )	自适应拒绝抽样法	n, $p(x)$, min, max, sp
rMARS( )	改进自适应拒绝抽样法	n, $p(x)$, min, max, sp, $infp$
rCCARS( )	凹凸分解自适应拒绝抽样法	n, $cvf(x)$, $ccf(x)$, min, max, sp
rASS( )	自适应切片抽样发	n, $init$, $p_0(x)$, w
rARMS( )	自适应Metropolis拒绝抽样法	n, $\ln p_0(x)$, $init$, $indFunc$

张东

R code

```
x <- rARS(10^5,"exp(-x^2/2)",-Inf,Inf,c(-2,2))
```


表1 rARS生成样本与真实分布特征统计量对比

统计量	25%分位数	中位数	均值	75%分位数	标准差
真实分布	-0.6745	0	0	0.6745	1
模拟样本	-0.6874	-0.0099	-0.0062	0.6708	1.0101

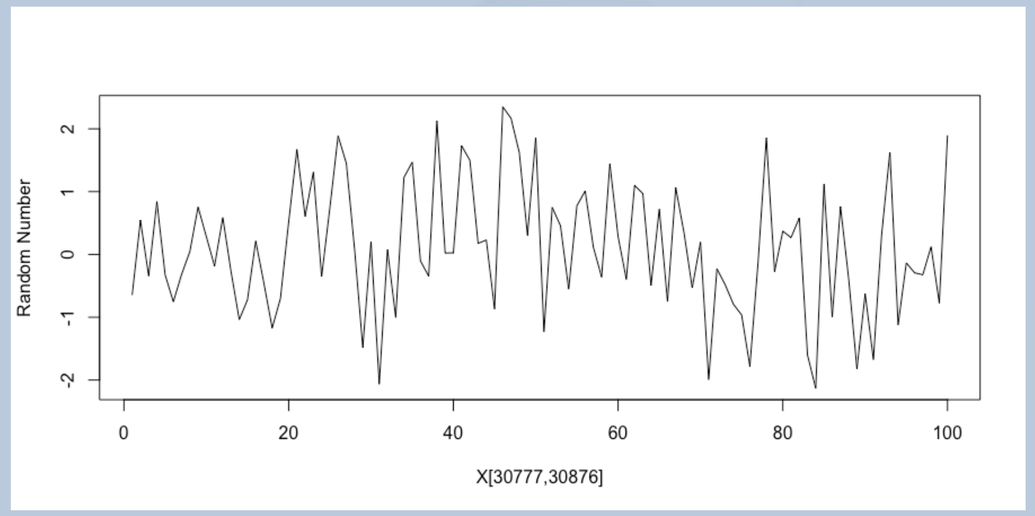


图7 rARS函数收敛图

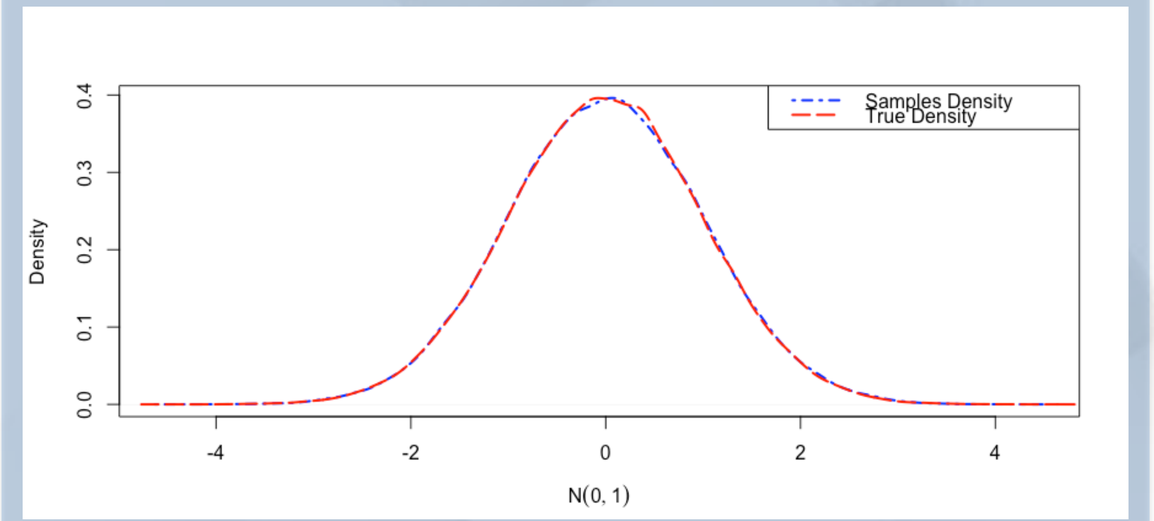


图8 rARS生成样本的密度估计与真实分布对比图

rMARS

R code

```
x <- rMARS(10^4,"exp(-(4-x^2)^2)",-Inf,Inf, c(-2.5,0,2.5),c(-2/sqrt(3),2/sqrt(3)))
```

表2 rMARS生成样本与真实分布特征统计量对比

统计量	25%分位数	均值	75%分位数	标准差
真实分布	-1.983	0	1.983	1.983
模拟样本	-1.984	0.001	1.982	1.982

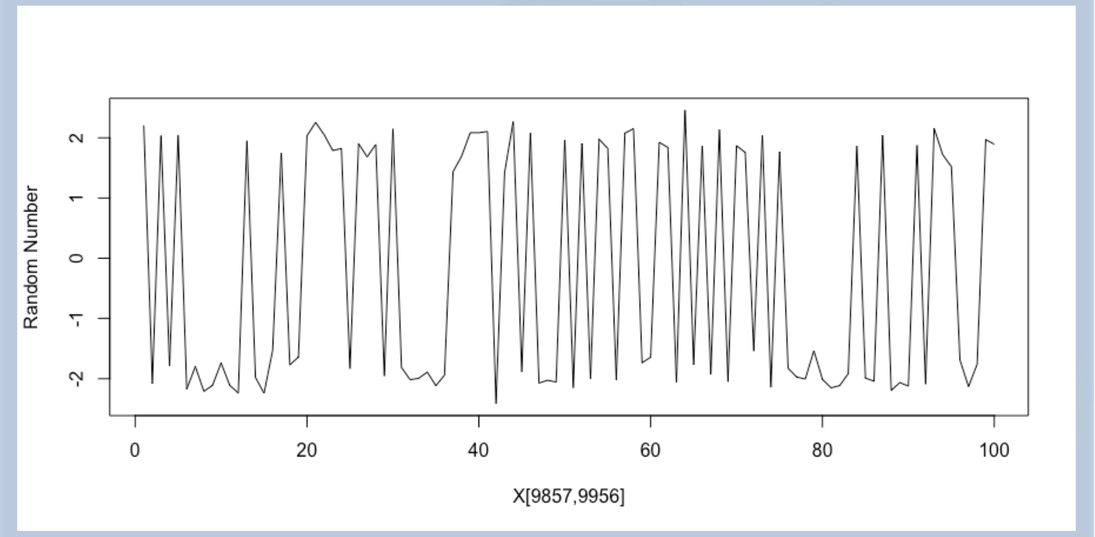


图9 rMARS函数收敛图

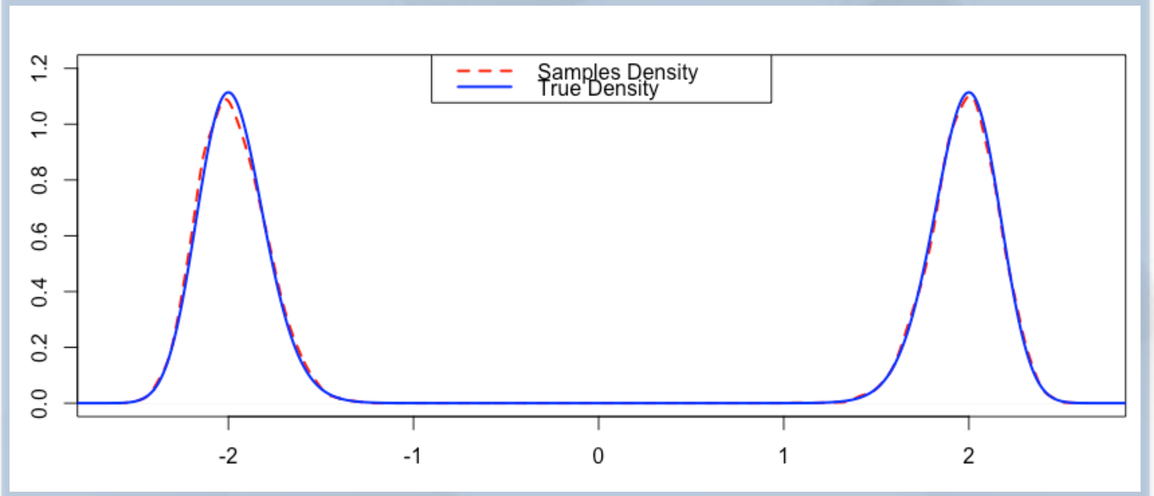


图10 rMARS生成样本的密度估计与真实分布对比图

R code

```
x <- rCCARS(10^4,"x^4", "-8*x^2+16",-3,3,c(-2,1))
```

表3 rCCARS生成样本与真实分布特征统计量对比

统计量	真实分布	模拟样本
25%分位数	-1.984	-1.9857
均值	0	-0.006
75%分位数	1.984	1.9799
标准差	1.9835	1.9798

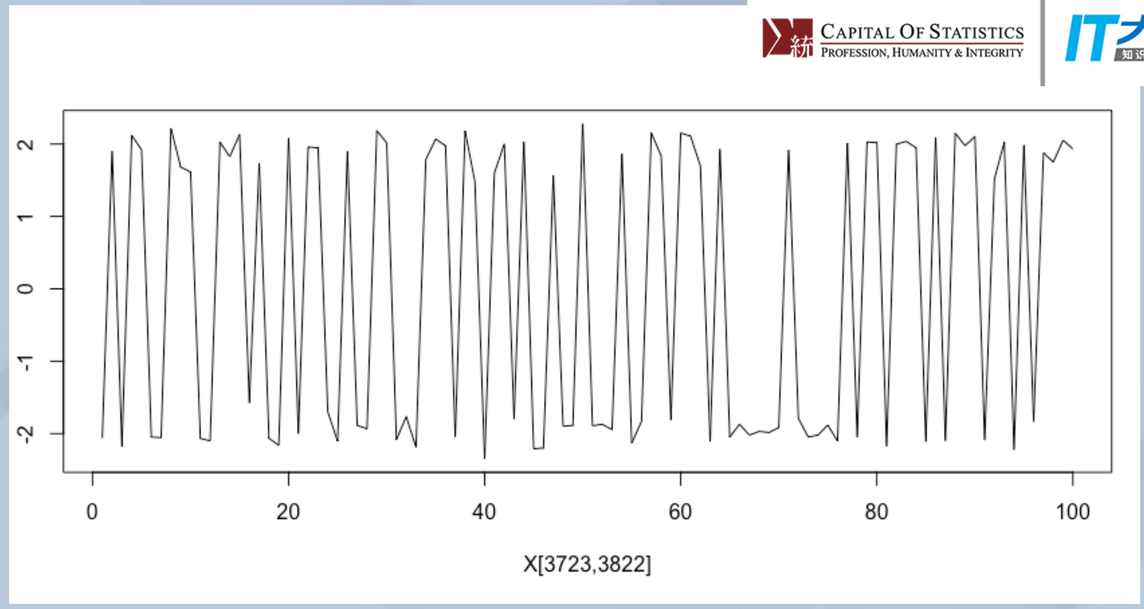


图11 rCCARS函数收敛图

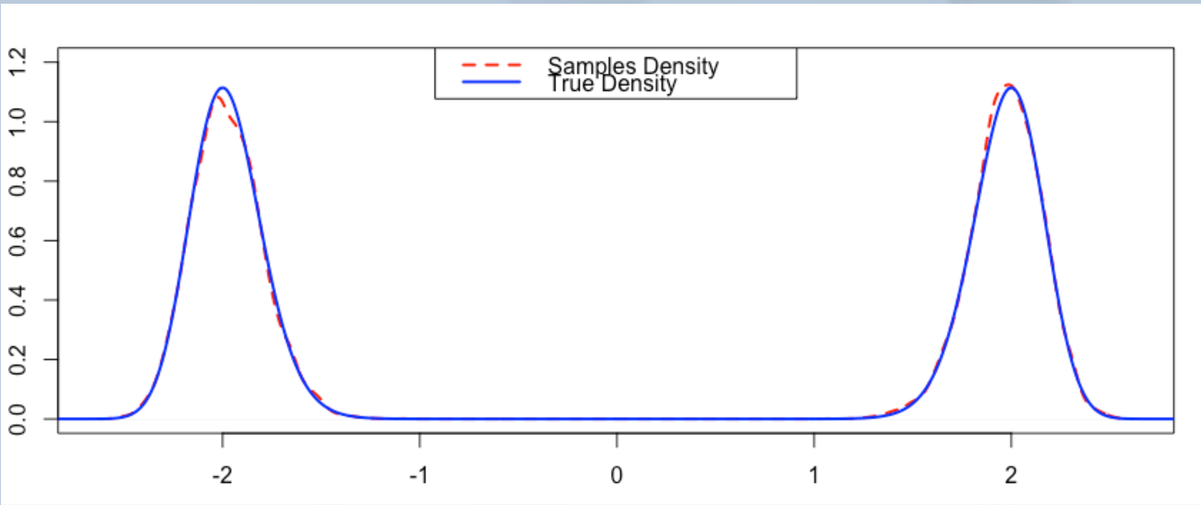


图12 rCCARS生成样本的密度估计与真实分布对比图

张东

R code

```
x <- rASS(10^4,-1,"1.114283*exp(-(4-x^2)^2)",3)
```

表4 rASS生成样本与真实分布特征统计量对比

统计量	真实分布	模拟样本
25%分位数	-1.983	-1.987
均值	0	-0.0136
75%分位数	1.983	1.987
标准差	1.9834	1.982

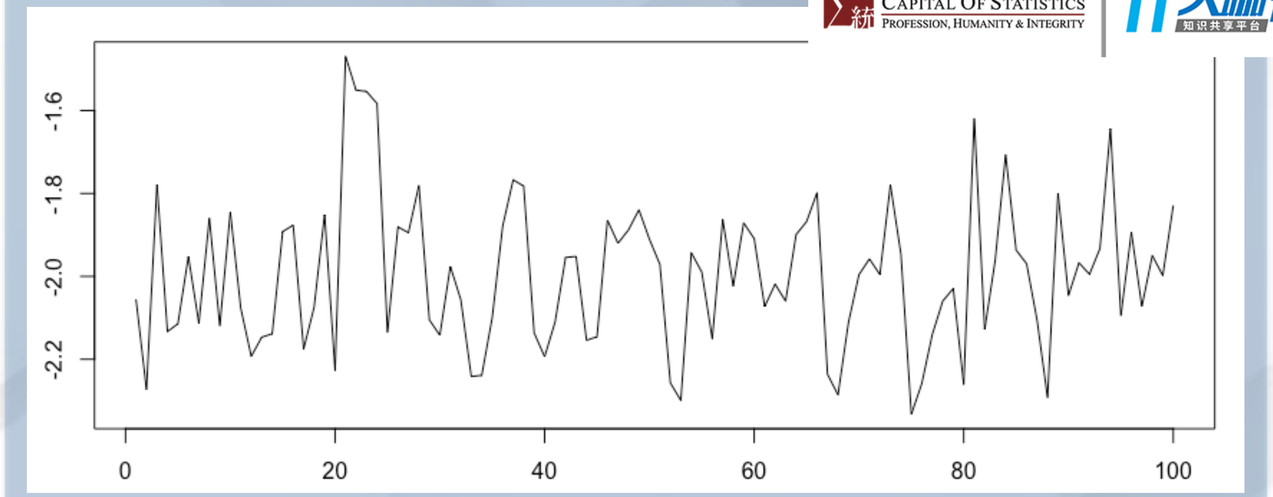


图13 rASS函数收敛图

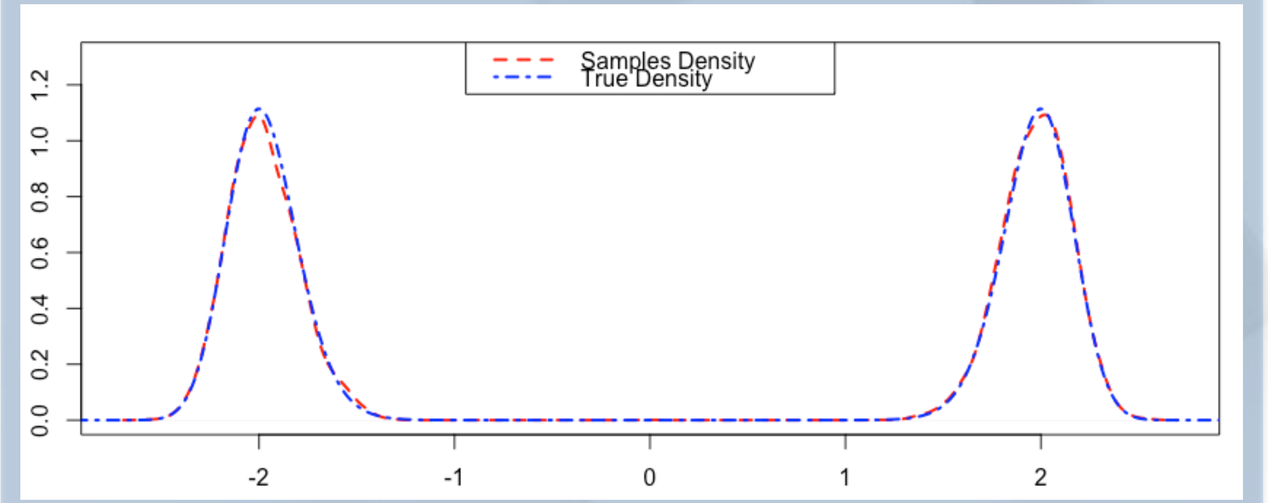


图14 rASS生成样本的密度估计与真实分布对比图

rARMS
HI PACKAGE

R code

```
p <- function(x) {1}
indFunc <- function(x) {(min(x)>0)*(max(x)<1)}
x <- rARMS(10000,c(0.1,0.9),p,indFunc)
```

表5 rARMS生成样本征统计量

维度	均值	标准差
1	0.495	0.291
2	0.511	0.286
真实分布	0.5	0.2887

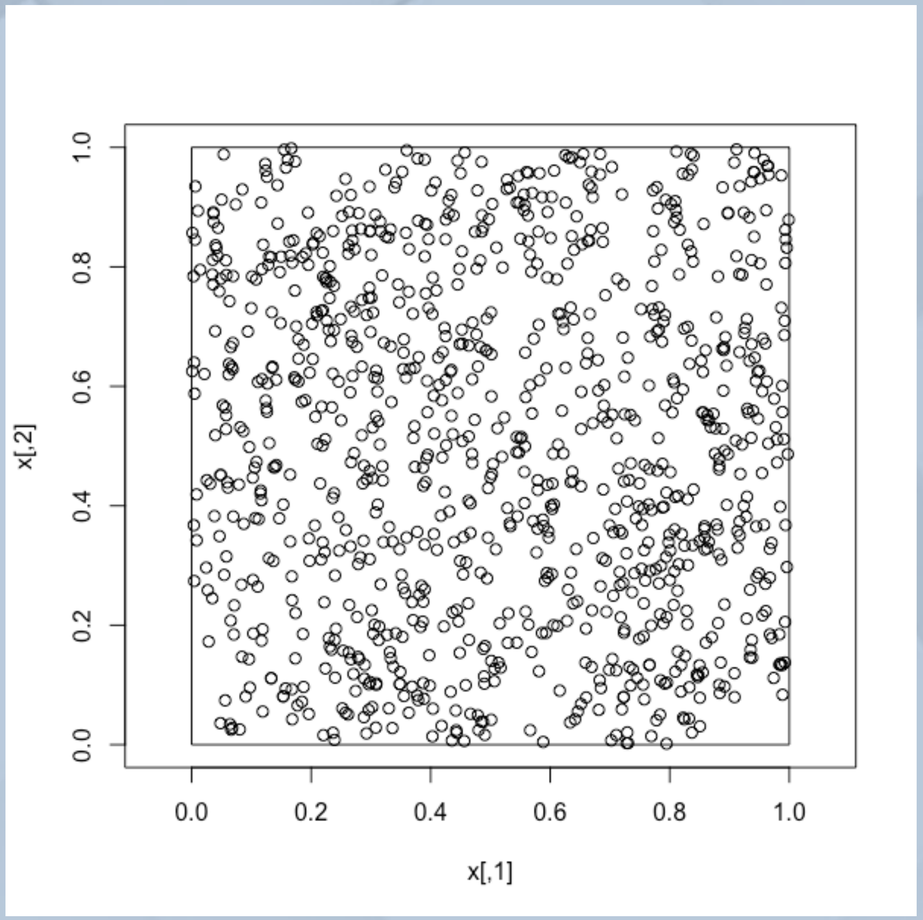


图15 rARMS生成二元均匀分布散点图

抽样结果 对比

混合正态分布 $p=0.3,\mu_1=2,\mu_2=8,\sigma_1=1,\sigma_2=8$，且$N$=10^4。

表6 AdapSamp包生成混合正态分布随机数的结果比对

	抽样结果		
抽样方法	均值	标准差	抽样效率
rMARS函数	6.246	3.271	1.7897
rCCARS函数	6.187	3.241	3.1501
rASS函数	6.039	3.312	1.5459
rARMS函数	6.130	3.291	1.2507

抽样效率：产生一个样本量的平均迭代次数。

A 抽样耗时过长

B 凹凸分解法的不唯一性会影响抽样效率

C 测试时间因平台不同而有所差异

A 压缩精简算法

B 对CCARS最优分解的效率研究

C 底层语言研究耗时

谢谢