arXiv:2305.02446v1 [stat.ME] 3 May 2023

利用局部枢轴方法提高精度：一种通用的方差缩减方法

Marcus Olofsson[1],Anton Grafstr¨om[2], Niklas L. P. Lundstr¨om[1]

[1]数学与数学统计学系，乌默奥大学，*SE-90187* 乌默奥，
瑞典 *niklas.lundstrom@umu.se*，
marcus.olofsson@umu.se

[2]森林资源管理系，瑞典农业科学大学，乌默奥，瑞典
anton.grafstrom@slu.se

2023年5月5日

摘要

局部枢轴法（LPM）是一种成功的抽样方法，可从离散总体中抽取分布均匀的样本。我们展示了如何利用 LPM 从任意连续分布中抽样，从而在一般情况下有效缩减方差。该方法可对任意维度的连续分布实现“自动分层”，并选取分布均匀的“稀疏”样本。我们通过多个示例展示 LPM 的简便性、通用性和有效性，包括积分的蒙特卡罗估计、期权定价，以及非线性动力系统中的稳定性估计。此外，我们还展示了如何将 LPM 与其他方差缩减技术（如重要性抽样）结合，以进一步缩减方差。为便于实现 LPM，我们提供了在 MATLAB 和 R 中使用 LPM 的快速入门指南，其中包含示例代码，演示如何仅用几行代码实现方差缩减。

关键词：方差缩减 ；空间抽样；分层抽样；不等概率抽样

1 引言

假设我们希望估计离散总体 $U = \{1, 2, ..., N\}$ 中某种特征的均值 μ。一种经典的估计量是 Horvitz-Thompson 估计量，参见 [13]

$$\hat{\mu}_{HT} := \hat{\mu} := \frac{1}{N} \sum_{i \in S} \frac{y_i}{\pi_i}, \tag{1}$$

其中， $S \subset U$ 是一个样本量为 n的样本， y_i 是个体 i的特征值，而 π_i 是在给定某种抽样设计下个体 i 的包含概率。对于相等的包含概率 $\pi_i = n/N$

这就相当于样本均值。随机样本 S 最好能使$\hat{\mu}$ 接近真实值 μ，并且在允许的样本集合中具有较小的方差。如果样本量可以任意大，且采用不放回抽样，那么只需取 $n = N$即可轻易实现这一点，从而得到真实值 μ，方差为 0。然而，如果收集或获取数值 y_i 的成本很高，就需要谨慎选择样本，以便估计量 (1) 即使在 n取值适中的情况下，也能给出可接受的结果并保持较小的方差。这正是采用方差缩减技术的动机。

方差缩减对于许多从事统计估计的领域而言都不可或缺。经典的方差缩减方法包括控制变量、对偶变量、分层抽样和重要性抽样，参见例如 [5]。分层抽样或许是最常用的方法之一，因为它简单且高效。其主要思路是将结果空间划分为较小的子区域，“层”，并在给定层内生成每个样本点，从而确保样本分布均匀且比例适当。这样，样本便能按设计覆盖整个总体。一般而言，已有充分证据表明，相较于采用独立观测进行抽样，此类分布均匀的样本效率更高，也更受青睐。分层抽样的一个缺点是，构造合适的层并允许入样概率不等可能并非易事，尤其是在辅助变量数量较多时。

基于 [4]，作者在 [10] 中提出了一种新颖的抽样方法，即局部枢轴法(LPM)，用于从离散总体中选取一个占总体比例很小且分布均匀的样本。概括而言，LPM 会自动进行分层，并从总体中选取一个“稀疏”且分布均匀的样本，同时允许入样概率可能不相等。在多种可获得多个辅助变量、且收集观测值的成本相对较高的应用中，LPM 已成功用于降低方差。需要昂贵的实地调查以测度目标变量的环境调查、森林和景观调查均采用 LPM。例如，瑞典国家森林清查和瑞典国家景观清查使用了这种抽样方法，见 [6] 和 [1]。LPM 还应用于农业调查（例如 [2], ）和土壤调查（例如 [3], ），并已用于从连续总体中抽样 [7, 8]。

尽管 LPM 灵活、易于实现（也适用于不等入样概率），并且已证明效率显著，但尚未被更广泛的群体采用。这正是本文的主要动机：我们将介绍 LPM 的基础知识，并展示其在任意维度的任意连续分布中的适用性。通过积分的蒙特卡罗估计、欧式期权定价，以及雨林和水力发电机转子稳定性的估计等示例，我们将展示如何以最少的工作量运用 LPM，在广泛的应用领域实现快速且可靠的方差缩减。

名为 `BalancedSampling` 的 R语言库（参见 [9], ）以及 MATLAB 实现均可直接获取。本文最后附有一份快速入门指南，介绍如何使用这些实现。

2 连续分布的 LPM

受林业应用的启发，LPM 最初旨在从离散总体中抽取分布均匀的样本。我们首先介绍这一过程。

令 U 表示一个由 N 个单元组成的离散总体，每个单元都有预先设定的入样概率 π_i，该概率可能由某个辅助可测变量 z_i 的函数给出。LPM 通过逐步更新总体的入样概率，从 U 中选取样本，使得在每一步中，至少有两个相邻单元中的一个的抽样结果得以确定。更具体地说，给定两个

nei相邻单元 i 和 j，我们随机更新它们的概率（π_i，π_j）t o

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) & \text{with probability } \frac{\pi_j}{\pi_i+\pi_j} \\ (\pi_i + \pi_j, 0) & \text{with probability } \frac{\pi_i}{\pi_i+\pi_j} \end{cases} \quad \text{如果 } \pi_i + \pi_j < 1,$$

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) & \text{with probability } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1) & \text{with probability } \frac{1-\pi_i}{2-\pi_i-\pi_j} \end{cases} \quad \text{if } \pi_i + \pi_j \geq 1. \tag{2}$$

换言之，我们将包含概率的概率质量从一个单元转移到另一个单元，使得接收概率质量的单元被纳入样本，或提供概率质量的单元被排除在样本之外。重复这一过程，直到所有单元的（更新后）包含概率均为 0 或 1，此时，相应的局部枢轴法样本 $S \subset U$由（更新后）包含概率为 1 的单元构成。该样本 S 在总体中分布均匀，其期望样本量为 $n = \sum_i \pi_i$。如果 n 为整数，则样本量是固定的。[1]

具体而言，LPM 在离散总体上的运作过程如下：

算法 1（用于选择样本的局部枢轴法）

$i)$ 随机选择 s 从概率更新后严格介于 *0* 和 *1* 的单元中选取 i

$.ii)$ 找到距离 j 最近的邻居 i。

$iii)$ 根据(2)更新单位 i 和 j 的概率。

$iv)$ 从 $i)$ 开始重复，直到所有单位的更新后概率都等于0或1。

注释 1 只需增加一些计算量，就能得到分布得更加均匀的样本，方法是将上述算法中的 $iii)$ 替换为

$iii')$ 如果 j 以 i 为最近邻，则根据(2)更新概率。

算法 *1* 中的过程通常足以满足实际应用的需求。在需要时，我们将使用 iii （iii'）的算法分别记为*“LPM2”*（*“LPM1”*）；但在一般讨论该方法时，我们简称其为 *LPM*。

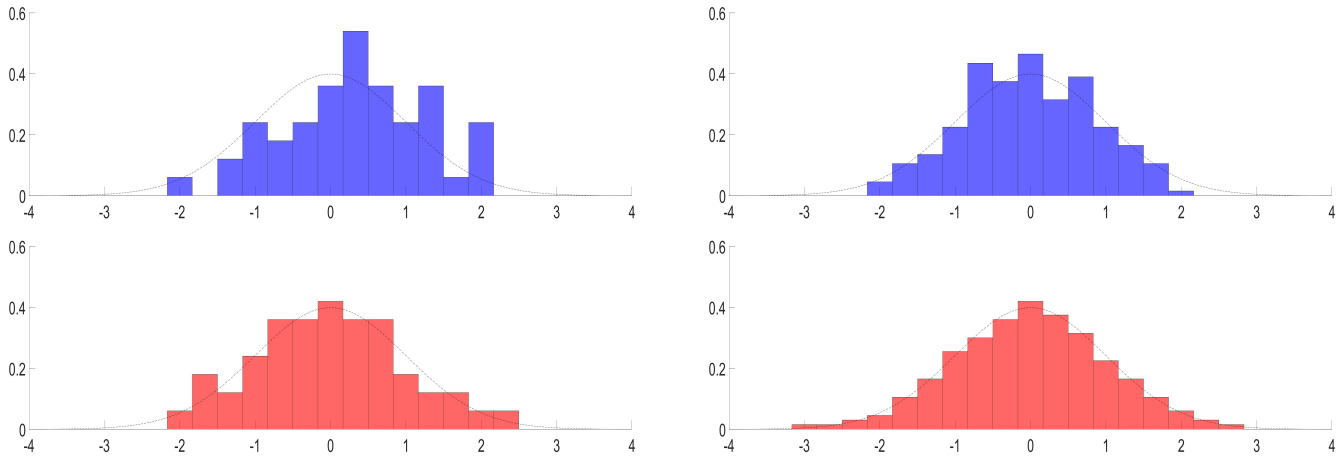
注 2 请注意，在步骤 $ii)$中确定“最近邻”时，可以使用任意距离函数 $d(i, j)$ 。特别是，距离可以在任意空间中度量，使用户能够在由大量变量构成的辅助空间中找到分布良好的样本。

2.1 扩展到连续分布

˜要将 LPM 应用于具有分布 U 的连续总体，我们只需在算法 1 中增加一个步骤——对总体进行离散化。

$\not{0})$ 从 U 中抽取 N 个独立点。

[1]这并非必 ry 为整数，但为便于阐述，我们将仅讨论这种情况 .



(a) $n = 50$ (b) $n = 200$

图 1：标准正态分布的独立同分布观测值（蓝色，顶部一行）和 LPM（红色，底部一行）样本的可视化比较。

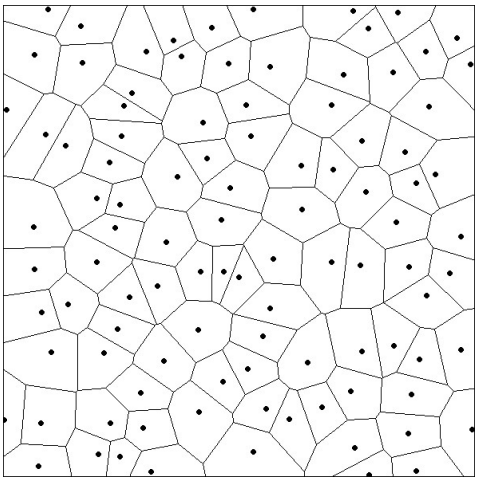
现在，我们可以对步骤 $\emptyset$ 中生成的离散子集应用算法 1 中的 *i)–iv)* 步骤，从而获得连续总体 U 的˜局部枢轴法样本。步骤 $\emptyset$) 中的离散化不会引入任何偏差。

将 LPM 应用于连续分布时，必须同时选择样本量 n 和离散化规模 N。对于 N 应该或需要多大，目前没有定论，但下面的示例表明，$N = 10 \cdot n$ 足以实现显著的方差缩减，而且 $N = 100 \cdot n$ 就足以获得 LPM 的大部分益处，见图 5。如果 $n = N$，那么 LPM 与独立同分布（iid）观测一致，无法实现方差缩减。提醒读者，性状评估只需针对 LPM 算法选出的 n 个点进行，因此通常可以以很低的成本增大 N 。

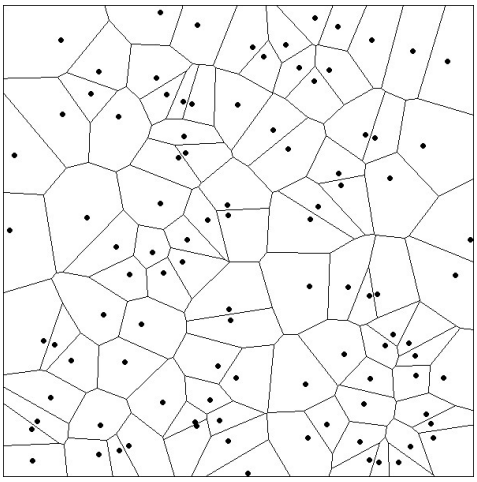
我们通过示例来展示 LPM 在连续分布上的简洁性和高效性。

示例 1（来自正态分布的均匀分散样本。） 我们从正态分布中抽取 $N = 10^4$ 个独立同分布点，然后对这个子集应用 *LPM2*。结果得到一个来自正态分布的均匀分散样本。图 *1a* 和图 *1b* 分别展示了 $n = 50$ 个点和 $n = 200$ 个点所生成的 *LPM*样本的直方图。为了进行直观比较，我们还提供了具有相同样本量的独立同分布观测值的直方图。

示例 2（来自二维均匀分布的分布均匀的样本。） 图*2*分别展示了包含 $n = 100$ 个观测值的 *LPM2* 样本以及来自标准 $2d$-均匀分布的 100 个独立同分布观测值的 *Voronoi* 多边形。空间平衡定义为 $\mathbb{E}\left[(1/n)\sum_{i=1}^{n}(a_i - 1)^2\right]$, 其中 a_i是 *Voronoi* 镶嵌中多边形 i 的面积，n 等于点的数量和总面积；基于使用 $n = 10^2$ 和 $N = 10^4$进行的 10^4次模拟，*LPM* 镶嵌的空间平衡约为 0.065 ，而独立同分布镶嵌的空间平衡约为 0.313 。



(a) LPM样本。



(b) 独立同分布数。

图2：沃罗诺伊多边形 分别为 100 个 LPM 和独立同分布的二维均匀分布随机数的多边形。

示例 3（蒙特卡罗积分估计。） 为了估计该积分

$$\mu := \int_0^1 x dx = \frac{1}{2}, \tag{3}$$

标准蒙特卡罗方法是从均匀分布 $(0, 1)$ 中抽取 n 个独立同分布观测值，并计算样本均值。这种方法之所以可行，是因为该积分可以表示为函数 $y(X) = X$ 的期望值，其中 X 服从 $U(0, 1)$ 分布，即 $\mu = \mathbb{E}[y(X)] = \int_0^1 y(x)f(x)dx$，其中 f 表示均匀概率密度函数，且对于 $x \in (0, 1)$，$f(x) = 1$，其他情况下为 0。在区间 $(0, 1)$ 上抽取 n 个均匀分布的观测值，会产生抽样强度函数 $\pi(x) = nf(x) = n$，其定义域为 $(0, 1)$。该估计量可以表示为

$$\hat{\mu} = \sum_{x \in S} \frac{y(x)}{\pi(x)} = \frac{1}{n} \sum_{x \in S} y(x),$$

其中，$S = \{X_1, X_2, \ldots, X_n\}$ 表示样本量为 n的随机样本。与使用独立观测相比，使用分布均匀的样本可以通过更小的样本量获得更好的估计量。

表1展示了使用由离散总体中10% 或1% 的样本点构成的 *LPM2* 样本时的数值结果，该总体的规模为 $N = 10^4$。使用10%的样本点时，*LPM* 可使标准差减半。这些结果与从10 个相等且均匀分布的层中抽取同等规模的样本所获得的结果相当。不过请注意，*LPM* 无需费力构建各层，而在更复杂的情况下，这可能需要相当大的工作量。图4a展示了在若干固定样本量 n下，估计量的方差如何随 N 变化而降低。

表1中 *LPM* 估计量和分层程序的标准差，是通过将每个估计量朴素地重复 $m = 10^4$ 次而得出的。关于如何根据单个样本估计 *LPM* 估计量的方差，请参阅第4节。

2.2 Com将 LPM 与其他方差缩减技术相结合

LPM 的一大优势在于其简单灵活；它可以直接应用于任何离散总体 U，以生成其稀疏版本，同时保留主要统计特征。这

	$\hat{\mu}_{iid}$	$\hat{\sigma}_{iid}$	$\hat{\mu}_{LPM}$	$\hat{\sigma}_{LPM}$	$\hat{\mu}_{strat}$	$\hat{\sigma}_{strat}$
$n = 10^2$	0.525	0.028	0.508	0.004	0.496	0.003
$n = 10^3$	0.500	0.009	0.502	0.002	0.499	0.001

表1：使用独立同分布的均匀随机数、LPM 和分层方法对 (3) 进行数值估计。LPM 的离散化使用 $N = 10^4$ 个点。

这使我们能够轻松地将 LPM 与其他方差缩减方法结合起来，尤其是在连续分布的情形下，我们可以将步骤 ∅）中的独立同分布离散化替换为基于方差缩减技术的离散化。更具体地说，在离散化连续分布 U 之前，我们可以使用任何合适的方差缩减技术，然后应用 LPM，从而同时获得两种技术带来的方差缩减效果。下面我们通过将 LPM 与重要性抽样结合来举例说明。

例4 假设我们希望估计 μ的期望值

$$y_{0.999}(X) = \begin{cases} 0 & if\ X \le \alpha_{0.999} \\ 1000X & if\ X > \alpha_{0.999}, \end{cases} \tag{4}$$

其中，$X \sim \mathcal{N}(0,\ 1)$，$\alpha_{0.999}$是 $\mathcal{N}(0,\ 1)$的99.9百分位数。可以通过大小为 N的独立同分布样本来实现，即 $\frac{1}{N}\sum_{i=1}^{N} 1000 X_i \mathbb{I}\{X_i > \alpha_{0.999}\}$，其中 X_i是第 i次抽样的值。然而，由于 $X_i > \alpha_{0.999}$ 是稀有事件，要获得较好的估计量，需要非常大的样本。

处理稀有事件时，一种常见的方差缩减技术是重要性抽样（*IS*）。其思路是改变标的资产的概率分布，使我们关注的事件更有可能发生，然后通过降低每个样本点的贡献权重来补偿这种改变。具体而言，在概率测度 $\mathbb{P}$ 下估计 $\mathbb{E}_{\mathbb{P}}[h(X)]$，其密度为 f；我们也可以转而在概率测度 $\mathbb{Q}$ 下考虑 $\mathbb{E}_{\mathbb{Q}}\left[h(X)f(X)/g(X)\right]$，其密度为 g，其中 $f(x)/g(x)$是 $\mathbb{P}$ 与 $\mathbb{Q}$ 之间的 *Radon–Nikodym* 导数。这是可行的，因为

$$\mathbb{E}[h(X)] = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}_{\mathbb{Q}}\left[h(X)\frac{f(X)}{g(X)}\right].$$

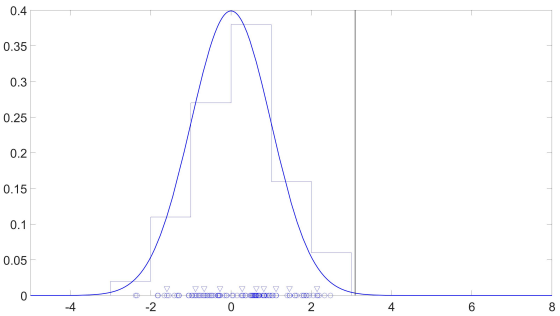
根据 g 的选择，估计后者可能会得到更小的方差。有关重要性抽样以及如何选择测度 $\mathbb{Q}$ 的更多信息，请参阅 *[5]* 。

更具体地说，若将 $\mathcal{N}(0, 1)$ 的密度记为 f，将 $\mathcal{N}(3, 1)$ 的密度记为 g，并利用重要性抽样估计 μ，则可得

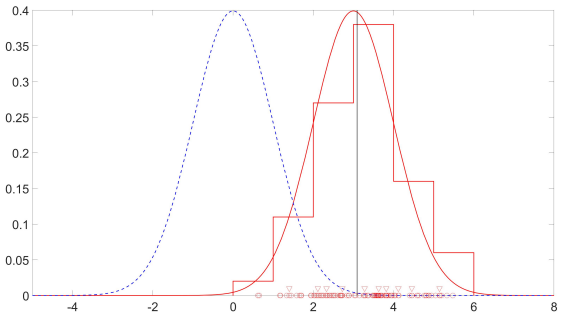
$$\mu = E_f[y_{0.99}(X)] = \int y_{0.99}(x)f(x)dx = \int y_{0.99}(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}_g\left[y_{0.99}(X)\frac{f(X)}{g(X)}\right]$$

其中，$\mathbb{E}_f$ 和 $\mathbb{E}_g$ 分别是在 f 和 g 下的期望值。从 $\mathcal{N}(3, 1)$ 中抽取独立同分布的观测样本，并利用这些样本估计最后一个表达式，可发现方差大幅缩减，最多可降低400 倍。随后可使用 *LPM* 对这些独立同分布样本进行稀疏化，使方差进一步降低10 倍，而不增加样本量。我们在表 *2* 中给出数值结果，并在图*3*中示意展示重要性抽样与 *LPM* 的结合方式。

在金融数学中，重要性抽样结合分层已被证明极为高效，见 *[5],* ，这一结果在 *LPM* 中也得到了明确验证。我们强调，*LPM* 无需任何调整，只需一行代码即可应用；与仅使用重要性抽样相比，只需使用$1\ 10$ 的样本点，便能获得近似相同的方差。



(a) 离散化分布 $\mathcal{N}(0,\ 1)$。



(b) 离散化分布 $\mathcal{N}(3,\ 1)$。

图3：平移离散化分布后，最终落在非零区域内的样本点显著增多。圆圈（○）表示离散化的独立同分布样本（$N = 100$），三角形（▽）表示 LPM 稀疏化（$n = 10$），阶梯线则是离散化样本的归一化直方图。黑色竖线表示 $\alpha_{0.999}$。

	$\hat{\mu}_{iid}$	$\hat{\sigma}_{iid}$	$\hat{\mu}_{LPM}$	$\hat{\sigma}_{LPM}$	$\hat{\mu}_{IS}$	$\hat{\sigma}_{IS}$	$\hat{\mu}_{IS\&LPM}$	$\hat{\sigma}_{IS\&LPM}$
$n = 10^2$	0	11.101	0	10.601	2.663	0.614	3.281	0.147
$n = 10^3$	6.424	3.352	3.094	2.043	3.332	0.191	3.412	0.063
$n = 10^4$	1.953	1.097	-	-	3.391	0.062	-	-

表 2：使用独立同分布、LPM2、重要性抽样（IS）以及重要性抽样与 LPM2 相结合的方法，对 $\mu = \mathbb{E}[y_{0.999}(X)] \approx 3.367$ 进行的数值估计。将 LPM 与重要性抽样相结合，可使估计量的方差降低 10 倍。LPM 的离散化采用 $N = 10^4$ 个点进行。

3 应用示例

LPM 显然可应用于连续分布的诸多领域；本节将介绍数理金融（期权定价）和动力系统（估计非局部稳定性）中的几个示例。

3.1 数理金融：欧洲期权定价

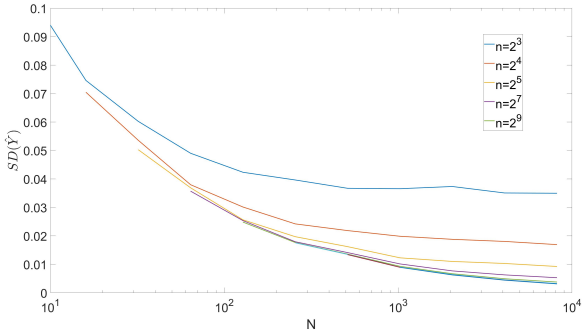
欧式看涨期权（欧式看涨期权）是一种金融衍生品，赋予持有人权利但没有义务在时间 $t = T$ 以价格 K 买入某项资产 S。资产 S 通常称为合约的标的资产，T 称为到期日，K 称为行权价。金融数学中的一个标准结论是，此类合约的公允价格等于其在所谓的风险中性测度下的折现期望值。事实上，将“公允价格”刻画为某一测度下的期望值，并非欧式期权所独有，而是普遍成立。因此，模拟和方差缩减是估算金融衍生品价格的常用工具。

在以 S 为标的资产、且其按几何布朗运动演化的欧式看涨期权这一特殊情形下，

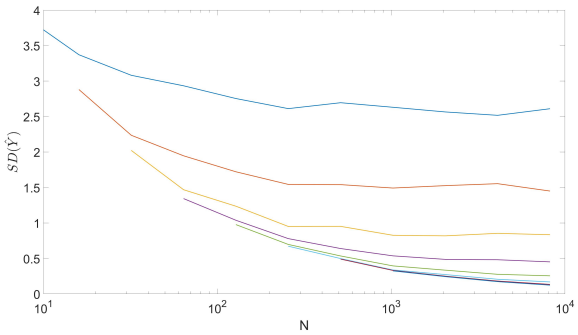
$$dS_t = \mu S_t dt + \sigma S_t dW_t, \qquad S_0 = s, \tag{5}$$

此价格由 $P_{EC} = \mathbb{E}\left[\max(0, S_T - K)\right]$ 给出，其中 $S_T = s\exp((r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z)$，$Z \sim \mathcal{N}(0,\ 1)$。在这一特定设置下，可以得到精确解，

$$P_{EC}(s) = N(d_1)s - N(d_2)Ke^{-rT}, \qquad \text{where} \tag{6}$$



(a) 示例3，$\mu := \int_0^1 x dx$。



(b) 示例应用3.1，欧式看涨期权。

图4：估计量的标准差随离散化参数 N 的变化情况（不同样本量 $n \leq N$）。

$$d_1 = \frac{1}{\sigma\sqrt{T-t}}\left[\log\left(\frac{s}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)\right] \qquad \text{and} \qquad d_2 = d_1 - \sigma\sqrt{T-t},$$

这使欧式看涨期权非常适合用于测试模拟技术和方差缩减方法。

这里考虑一份欧式看涨期权，参数为 $s = 100$、$r = 0.03$、$\sigma = \frac{1}{2}$、$T = 1/4$和 $K = 120$，其由式(6)给出的理论价格为 $P_{EC} \approx 3.886$。表3列出了通过模拟并使用和不使用LPM进行方差缩减所得到的价格估计。使用LPM可显著缩减方差。

方差$\hat{\sigma}_{LPM}$通过重复实验 $m = 10^4$次进行朴素计算，而$\hat{V}(\hat{P}_{EC}^{LPM})$则按照第4节所述，使用10个最近邻进行计算。图4b显示，对于若干固定样本量 n，估计量的方差如何随离散化参数 N 增大而下降。

	$\hat{P}_{EC}^{iid}$	$\hat{\sigma}_{iid}$	$\hat{P}_{EC}^{LPM}$	$\hat{\sigma}_{LPM}$	$\sqrt{\hat{V}_{LPM}(\hat{P}_{EC}^{LPM})}$
$n = 10^2$	3.523	0.899	4.089	0.307	0.374
$n = 10^3$	3.748	0.365	3.899	0.116	0.043

表3：使用独立同分布正态随机数和 LPM2样本得到的 $P_{EC} \approx 3.886$ 估计量。LPM 的离散化采用 $N = 10^4$ 个点。最后一列为第4节中的方差估计。

3.2 非线性动力学：估计非局部稳定性

动力系统（DS）是一组描述点随时间变化的微分方程。例如，坠落物体的运动、时钟摆的摆动、水在管道中的流动，以及湖中鱼的数量变化。动力系统可以写成如下形式

$$\frac{dx}{dt} = f(x,t), \qquad x = x(t) \in \mathbf{R}^d, \quad t > 0, \tag{7}$$

其中，$f(x,\ t)$ 是定义在 $\mathbf{R}^d \times \mathbf{R}$ 上、取值于 $\mathbf{R}^d$ 的给定函数，d 是空间维数。初始时，在 $t = 0$，x 由一个初始条件给定

$$x(0) = (x_1(0), x_2(0), \ldots, x_d(0)) \in \mathbf{R}^d.$$

了解动力系统解的稳定性对许多应用都非常重要。量化稳定性的一种简单方法是检验解抵御扰动的能力。对于非线性动力系统，这类研究自然分为局部分析和非局部分析。局部稳定性方法考察小扰动，通常基于线性化，只能提供解在小邻域内的信息。我们在此考虑的非局部方法也考察大扰动，因此需要研究动力系统解的吸引域，例如参见 [21, 17, 14]。事实上，我们可以估计运行中的水力发电机（如下方第二个示例所示）的（非局部）稳定性：首先将其建模为动力系统，然后求解该动力系统，从而找到代表运行状态的解，最后反复检验动力系统在受到给定扰动后是否能恢复到同一解。如果对大量扰动都能恢复该解，我们就称其稳定。扰动可以自然地通过系统的初始条件来建模。因此，要了解动力系统的稳定性，就需要从一组不同的初始条件出发求解方程，并研究由此产生的行为。如果某个初始条件能使系统在合理的时间限制内达到预期解，我们就称其安全，否则则称其不安全。令 $\mathbb{N}_{\text{safe}}$ 和 $\mathbb{N}_{\text{tot}}$ 分别表示安全初始条件的数量和经过测试的初始条件数量。我们通过以下方式得到一个简单的稳定性度量：

$$\mathcal{P} = \frac{\mathbb{N}_{\text{safe}}}{\mathbb{N}_{\text{tot}}} \tag{8}$$

即安全初始条件所占的比例。测度 $\mathcal{P}$ 已在例如 [15, 17, 14] 中得到研究；关于非局部稳定性度量的进一步讨论和应用，我们也请读者参阅该文。接下来需要选择一组合适的初始条件（扰动），用于检验式 (7) 中的动力系统。这些扰动可以从预先设定的概率分布中确定性地选取，也可以随机抽取；选择时应考虑系统在现实中可能遭遇的情况，因此需视具体情况而定。

由于方程组 (7) 可能需要大量计算资源且耗时较长，因此在随机抽取初始条件时实施方差缩减是有益的。还要指出，除了稳定性之外，与之相关的韧性概念也适用于方差缩减，构造和应用非局部韧性度量的示例可参见 [18, 16, 14] 。下面我们介绍两个示例：第一个源自 [17]，研究雨林；第二个拓展了 [15] 的研究，探讨发电机中的非局部稳定性。

3.2.1 雨林的一个简单模型

亚马逊雨林可视为具有两种稳定状态：肥沃森林状态和贫瘠稀树草原状态。亚马逊地区的这种双重稳定性（也称为双稳态）源于雨林中的一个正反馈环：深根树木吸收储存在土壤中的水分，并将其蒸腾到大气中。因此，森林覆盖区域的总体降水量会增加；如果某个相当干旱的区域的森林覆盖超过一定的临界阈值，它仍可能适宜森林生长。另一方面，如果森林覆盖低于这一阈值，该区域的树木就会全部消失。这种行为可用以下简单的动力系统概括：

$$\frac{dx}{dt} = F(x) - Mx, \quad \text{where} \quad F(x) = \begin{cases} Rx(1-x) & \text{if} \quad x > x_{crit}, \\ 0 & \text{if} \quad x \le x_{crit}. \end{cases}$$

这里， $x = x(t)$ 是相对森林覆盖率， R 表示增长率， M 表示死亡率， x_{crit} 为临界森林覆盖率阈值。该模型的两个平衡态分别是森林态 x_F 和稀树草原

态 x_S，其表达式为

$$x_F = 1 - \frac{M}{R} \quad \text{and} \quad x_S = 0,$$

分别。当 $x_F > x_{crit}$ 时，两个平衡态均存在且稳定。

我们通过施加与森林生物量呈正态分布的下降相对应的扰动来检验模型，因此从 $\mathcal{N}(0,\ 1)$ 分布中抽取 n 个点 $\{z^1, z^2, \ldots, z^n\}$，并生成 n 个不同的初始条件，如下所示

$$x^i(0) = x_F - |z^i|, \quad i = 1, 2, \ldots, n.$$

为从初始条件出发对轨迹进行数值积分，我们使用 MATLAB 的 ODE 求解器 ODE45，并采用标准容差设置。我们对每条轨迹进行积分，直到解轨迹 $x(t)$ 到达以下两个邻域之一

$$(x_F - \epsilon, x_F + \epsilon) \quad \text{or} \quad (x_S - \epsilon, x_S + \epsilon),$$

re它们分别表示森林平衡态和稀树草原平衡态 .

在图 5a 中，我们展示了非局部稳定性度量 $\mathcal{P}$ 的模拟结果；该度量统计了安全初始条件所占的比例，定义见式 (8)。结果同时显示了使用和未使用 LPM 进行方差缩减的情况。对于每个 x_{crit} 值，我们使用 LPM 抽取了 $n = 50$ 个初始条件，并在正态分布的初始独立同分布离散化中使用了 $N = 10^4$ 个点。可以清楚地看到，方差显著缩减。我们还看到，随着临界阈值 x_{crit} 接近 0.5 处的分岔值，稳定性度量 $\mathcal{P}$ 迅速下降。这意味着，尽管森林能够存续（也就是说，x_F 存在），但这种情况极不可能发生，因为现实中很可能存在微小扰动。

3.2.2 一台电动水力发电机

大型同步发电机的转子与定子之间通常有很小的气隙。通常，气隙约为定子半径的 0.2%。对发电机的测量表明，由于生产缺陷等原因，气隙存在不对称性。这些不对称性会使气隙中的磁通密度发生畸变，从而在转子与定子之间产生吸引力，通常称为不平衡磁拉力（UMP）。不平衡磁拉力可能引发振动，最终可能危及机器安全。以下动力系统是 [15] 中推导出的水力发电机运动方程的无量纲形式:

$$\begin{aligned} X'' + 2\zeta X' + X &= F_X(X, Y, \tau), \\ Y'' + 2\zeta Y' + Y &= F_Y(X, Y, \tau). \end{aligned} \tag{9}$$

这里， X 和 Y 表示转子中心的位置， $\zeta = 0.1$为阻尼比， τ 为无量纲时间。根据物理考量，可得出力 F_X 和 F_Y 如下:

$$\begin{aligned} F_X &= \frac{k_m}{2\pi k} \int_0^{2\pi} \frac{\cos\varphi}{G(X, Y, \tau, \varphi)^2}\, d\varphi, \\ F_Y &= \frac{k_m}{2\pi k} \int_0^{2\pi} \frac{\sin\varphi}{G(X, Y, \tau, \varphi)^2}\, d\varphi, \end{aligned} \tag{10}$$

其中， $k \approx 3.456 \cdot 10^8 N/m$为机械刚度， $k_m \approx 1.4715 \cdot 10^8 N/m$为电磁刚度，气隙由下式给出

$$G = 1 + (\Delta - X)\cos\varphi - Y \sin\varphi. \tag{11}$$

参数 Δ 用于模拟发电机的缺陷：机器静止时，转子中心与定子中心相距 Δ（这会导致气隙不对称，进而产生 UMP）。这种不对称现象已在发电机领域得到广泛研究，称为偏心。

将速度 X' 和 Y' 作为额外的因变量，可以把式（9）中的二阶二维动力系统转化为式（7）形式的四维动力系统，用于描述转子在（$X(\tau)$ $Y(\tau)$ $X'(\tau)$ $Y'(\tau)$）中的运动，其中 $\tau > 0$。只要偏心 Δ 足够小，系统（9）就具有稳定平衡点 E_Δ。当 $\Delta = 0$ 时，机器处于理想状态，且 $E_0 = (0, 0, 0, 0)$。随着 Δ 增大， E_Δ 逐渐偏离原点，气隙变得不对称。当 Δ达到某个特定值时，平衡点 E_Δ 会在所谓的折叠分岔中消失，见 [15]。除平衡点 E_Δ外，当气隙变为非正值时，系统（9）始终存在转子—定子接触状态，这对应于机器彻底失效。

对于所研究的发电机，自然可以将转子中心的位移、施加在转子上的速度冲击，或两者的组合视为扰动。速度冲击可能由转子受到冲击引起，而位移则可能由机器底座受到冲击引起。我们选择测试机器在位移和速度均服从正态分布的扰动下的表现，从而得到一个四维扰动空间（ X 和 Y 方向上的位移，以及 X和 Y方向上的速度）。我们从以平衡点 E_Δ 为中心、四个方向上的标准差均为 $\sigma = 0.4$ 的四维正态分布中抽取 n 个点。具体来说，我们抽取 n 个点

$$\{z^1, z^2, \cdots, z^n\} \quad \text{where} \quad z^i = (z_1^i, z_2^i, z_3^i, z_4^i) \sim \mathcal{N}^4(E_\Delta, \sigma)$$

并将我们的 n 初始条件定义为

$$(X^i(0), Y^i(0), X'^i(0), Y'^i(0)) = E_\Delta + (z_1^i, |z_2^i|, z_3^i, z_4^i), \quad \text{for} \quad i = 1, 2, \ldots, n.$$

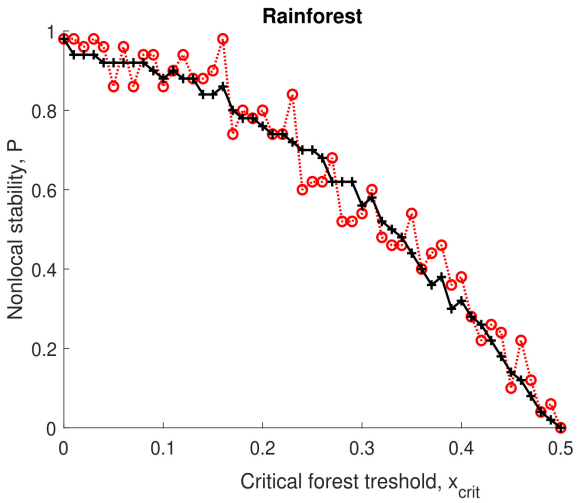
这里，对 z_2^i 取模将样本空间分为两半。由于发电机模型具有对称性，这是可行的；我们只需考虑半空间 $Y \geq 0$，因此将初始条件镜像映射到该集合中。为了从初始条件出发对轨迹进行数值积分，本文使用 MATLAB 的 ODE 求解器 ODE45，并采用标准容差设置。我们对每条轨迹进行积分，直到其到达以 E_Δ为中心、半径为 0.01 的小邻域（即一个球），或者直到气隙首次变为非正值；两者分别代表稳定运行和机器完全故障。

在图 5b 中，我们展示了非局部稳定性度量 $\mathcal{P}$ 的模拟结果；该度量按照公式 (8) 的定义，统计安全初始条件所占的比例，并比较了使用和不使用 LPM 进行方差缩减的情况。对于每个偏心 Δ 值，我们使用包含 $n = 500$ 个初始条件的 LPM 样本，并在初始独立同分布离散化中使用 $N = 10^4$ 个点进行模拟。与上文讨论的雨林模型一样，可以观察到方差有所降低。

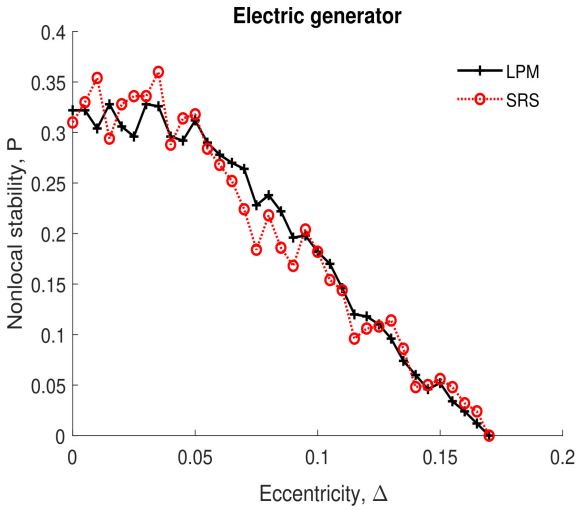
我们指出，公式 (10) 中的力 F_X 和 F_Y 以积分形式给出，在对动力系统 (9) 进行数值求解时，必须在每个时间步对其进行计算。这使得求解过程非常耗时。因此，在随机推断扰动时，采用本文提出的稳健方差缩减方法显然是有充分理由的，因为它能够显著降低稳定性研究的成本。

4 方差估计

采用许多方差缩减技术时，降低方差的估计量往往难以同时估计其方差。最朴素的估计量 $\hat{}\mu_{iid}$ 对 $\mu = \mathbb{E}(y(X))$ 的方差，



(a) 测度 $\mathcal{P}$ 随临界森林阈值 x_{crit} 变化的函数关系，分别展示采用和未采用方差缩减的情况。对于 x_{crit} 的每个取值，均考察 50 组初始条件。参数设置为 $R = 1$ 和 $M = 1/2$。



(b) 测度 $\mathcal{P}$ 随偏心率 Δ 的变化情况，以及采用和未采用方差缩减方法时的对比。对于 Δ 的每个取值，均考察了 500 个初始条件。每个数据点均需耗费数天时间才能完成。

图 5：雨林模型中森林平衡态（左）和发电机模型中运行平衡态（右）的稳定性。

基于随机变量 X的 n 个独立观测，可以通过样本方差轻松估计为

$$\hat{V}(\hat{\mu}_{iid}) = \frac{1}{n(n-1)} \sum_{x \in S} \left(y(x) - \frac{1}{n} \sum_{x' \in S} y(x') \right)^2 .$$

应用方差缩减技术时，方差的估计通常会更加复杂。对于 LPM，我们建议采用局部均值方差估计量来估计方差，如 [19, 20] 和 [12], 所示，另请参见 [7]。更确切地说，我们建议使用局部均值方差估计量来估计 $\mathbb{E}(y(X))$ 的 LPM 估计量的方差。

$$\hat{V}(\hat{\mu}_{LPM}) = \frac{n'}{n^2(n'-1)} \sum_{x \in S} \left(y(x) - \frac{1}{n'} \sum_{x' \in S_x} y(x') \right)^2 . \tag{12}$$

其中， n' 是点 x的局部邻域 $S_x \subset S$ 的大小，该邻域由点 x 及其 $n'-1$个最近邻组成。注意，当 $n' = n$时，估计量$\hat{}\sigma^2_{LM}$ 与$\hat{}\sigma^2_{iid}$完全相同。显然，估计量（12）取决于 n'，因此需要谨慎选择该参数。最优 n' 以及所提出的方差估计量的性能需要进行全面研究，但这超出了本文的研究范围；我们仅请读者参考表3中的估计结果，这些结果是在 $n' = 10$条件下计算得到的。

5 快速入门教程

为了方便读者　　因此，我们在此提供一份快速入门指南，介绍如何在 R 和 MATLA 中使用 LPM　　B.

5.1 R

LPM 在 R 中有两个不同版本的实现，分别为 `lpm1` 和 `lpm2`，两者均使用欧几里得距离。这些实现收录于软件包 `BalancedSampling` 中。语法如下:

`s=lpm1(prob,X)` or `s=lpm2(prob,X)`

其中

`prob` - 具有入样概率`X` 的长度为 N 的向量 $N \times q$ - 表示离
散总体的`s` 矩阵 X - 构成平衡局部枢轴法样本的行号。

一个具体示例是生成图2 a)，如下方代码所示。需要强调的是，与生成图2 b)相比，唯一的区别在第 6 行，即“select LPM-sample”。

```
library(BalancedSampling); library(deldir)      # 导入库
set.seed(1);                                    # 设置随机种子和样本大小
N = 10000; n = 100;                             # 设置样本量
p = rep(n/N,N);                                 # 集合（相等）纳入概率
X = cbind(runif(N),runif(N));                   # 离散化总体
s = lpm2(p,X); X=X[s,];                         #选择局部枢轴法样本
LPMtesselation = deldir(X[1:n,])                # 创建镶嵌（deldir 包）
LPMtiles = tile.list(LPMtesselation)
plot(LPMtiles, pch=19)                          # 绘制图形
```

5.2 MATLAB

在 MATLAB 中，提供了采用任意距离函数的 LPM2 实现及其方差估计量。为提高计算速度，此实现较为耗费内存，最适用于规模为 $N = 10^4$ 或更小的离散总体。语法如下:

`[s,svar]=lpm2(prob,X,distfcn,ns)`

其中

`prob` - 长度为 N、包含入样概率的向量

`X` - 表示离散总体的 $N \times q$ 矩阵

`distfcn` - 距离函数。MATLAB 标准距离函数均可使用，例如 'euclidean'、'cityblock' 或 'chebychev'。也可以使用用户指定的距离函数，参见备注3。

`ns` - 一个正整数，表示用于方差估计的最近邻数量（包括该点本身）。默认值为 `ns=1` ，此时将忽略方差估计。

`s` - 构成平衡局部枢轴法样本的 X 行号。

`svar`- $M \times (ns-1)$ 矩阵，其中 $M = length(s)$，第 i 行中的元素是 X 的行号，这些行表示与 `s(i)` 最近的最近邻（它们也位于 `s`中）。

备注3 *MATLAB* 函数 *lpm2*中允许使用任何用户指定的函数；此时， *distfcn* 应为字符串 *@fcnhandle*，其中 *fcnhandle* 是一个用户指定的函数，它接受以下参数：一个包含 X单行的~ $1 \times d$向量 X_i，一个包含 X多行的 $M \times d$矩阵 X，并返回一个距离向量 $M \times 1$，其第 j个元素是观测值~X_i 与 $X(j,:)$之间的距离。

MATLAB 中的一个具体示例是计算表3中的$\hat{P}_{EC}^{LPM}$，具体由以下代码完成。我们再次强调，方差缩减LPM估计与基于独立同分布观测值的估计之间唯一的区别，就是一行代码，即第5行"选择局部枢轴法样本"。

```
rng(1); N = 10000; n = 100; % 设置随机种子和样本量p=ones(N,1)*n/N; % 设置（相等的）纳入概率S0=100; K=120; T=1/4; r=.03; sigma=1/2; % 问题参数
X=randn(N,1); % 离散化总体s=lpm2(p,X,'euclidean'); X=X(s); p=p(s); %选择局部枢轴法样本stockvalue= S0*exp((r-1/2*sigma^2)*T + sigma*sqrt(T).*X(1:n));
optionpayoff = exp(-r*T)*max( 0, stockvalue - K ); % 股票和期权价值
priceest=sum(optionpayoff./p(1:n))/N %最终价格估计
```

参考文献

[1] Allard, A., Keskitalo, E.C.H., & Brown, A.（编）。（2023）。生物多样性监测：结合环境与社会数据（第1版）。Routledge。[2] Benedetti, R., Piersimoni, F., & Postiglione, P.（2015）。农业调查的空间单元抽样。柏林：Springer。[3] Brus, D. J.（2021）。空间抽样调查的统计方法：持续存在的误解与新进展。《欧洲土壤科学杂志》，72(2)，686-703。[4] Deville, J-C, Tille, Y.（1998）通过分割法进行不等概率无放回抽样，《生物计量学》85(1)，89-101。[5] Glasserman P.，金融工程中的蒙特卡罗方法，第53卷。纽约：Springer，2004。[6] Grafstr¨om, A., Zhao, X., Nylander, M., & Petersson, H.（2017）。一种应用于瑞典国家森林清查临时群组的森林清查新抽样策略。《加拿大森林研究杂志》，47(9)，1161-1167。[7] Grafstr¨om, A., & Matei, A.（2018）。连续总体的空间平衡抽样。斯堪的纳维亚统计学杂志，45(3)，792-805。[8] Grafstr¨om, A., Schnell, S., Saarela, S., Hubbell, S. P., & Condit, R.（2017）。森林清查的连续总体方法及其在设计中对信息的运用。*Environmetrics*，28(8)，e2480。

[9]Grafstr¨om, A., Lisic, J., Prentius, W.（2022）。BalancedSampling：均衡抽样与空间均衡抽样。R 包版本 1.6.3，网址 https://CRAN.R-project.org/package=BalancedSampling.[10] Grafstr¨om A., Lundstr¨om N.L.P., Schelin L.，通过关键方法实现空间均衡抽样，《Biometrics》68.2（2012）：514–520.[11] Grafstr¨om A., Lundstr¨om N.L.P.，为什么充分分散的概率样本是均衡的，《开放统计学期刊》3.1（2013）：36–41.[12] Grafstr¨om, A., & Schelin, L.（2014）。如何选择代表性样本。《斯堪的纳维亚统计学杂志》，41(2)，277-290。[13] Horvitz, D. G., & Thompson, D. J.（1952）。有限总体中不放回抽样的一般化。美国统计学会会刊，47(260)，663-685。[14] Lundstr¨om N.L.P.，如何寻找简单非局部稳定性和韧性度量，《非线性动力学》93.2（2018）：887–908。[15] Lundstr¨om N.L.P., Aidanp¨a¨a.，发电机形状偏差引起的电磁拉力的动态后果，《声音与振动》期刊 301.1–2（2007）：207–225。[16] Lundstr¨om N.L.P., Loeuille N., Meng X., Bodin M., Br¨annstr¨om, ̊A. 通过同时捕捞幼体和成体实现产量与保护目标，《美国博物学家》，193(3)，（2019）：373–390。[17] Menck P.J., Heitzig J., Marwan N., Kurths, J.），吸引域稳定性如何补充线性稳定性范式，《自然·物理》，9(2)，（2013）：89–92。[18] Mitra C., Kurths J., Donner R. V.，一种基于生态韧性的复杂系统多稳态综合量化指标，《科学报告》，5(1)，（2015）：1–10。[19] Stevens Jr, D. L., & Olsen, A. R.（2003）。环境资源空间均衡样本的方差估计。《Environmetrics》，14(6)，593-610。[20] Stevens Jr, D. L., & Olsen, A. R.（2004）。自然资源的空间均衡抽样。美国统计学会会刊，99(465)，262-278。[21] Tan Y., Ne˘si´c D., Mareels I.，关于极值搜索控制的非局部稳定性性质。《Automatica》42.6（2006）：889–903。