Stochastik-Praktikum

Zufallszahlen und Monte Carlo

Peter Frentrup

Humboldt-Universität zu Berlin

17. Oktober 2017

Übersicht

1. Erzeugung von Zufallszahlen
   - Erzeugung Bernoulli–verteilter Zufallszahlen
   - Die Inversionsmethode für univariate Verteilungen
   - Erzeugung multivariat normalverteilter Zufallszahlen

2. Monte-Carlo-Methoden
   - Monte Carlo-Integration
   - Crude Monte Carlo Integration
   - Monte-Carlo-Bänder

Johann von Neumann:

„Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.“

Definition

Ein Generator (uniformer) Pseudozufallszahlen ist ein Algorithmus, der von einem Startwert u_0 (seed) und einer Transformation T ausgehend, eine rekursive deterministische Zahlenfolge $u_n = T^{\circ n}(u_0)$ ([0,1]–wertiger) Folgeglieder erzeugt, die sich wie eine zufällige i. i. d. Folge von echten (uniformen) Zufallszahlen verhalten soll.
Monte Carlo Methoden basieren auf der häufigen Wiederholung eines Zufallsexperimentes bzw. Erzeugung von Pseudozufallszahlen um (zumeist komplexe) analytische Probleme näherungsweise zu lösen, wobei das Gesetz der großen Zahlen die Grundlage hierfür bildet.

Pseudo- und Quasi-Zufallszahlen

Pseudo-Zufallszahlen:
Eine Zahlenfolge, deren Bildungsgesetz möglichst schwer zu „erraten" ist.

Quasi-Zufallszahlen:
Eine Zahlenfolge, deren Häufigkeitsverteilung gemäß eines vorgegebenen Abstandsbegriffs einer vorgegebenen Wahrscheinlichkeitsverteilung (i. d. R. $\mathcal{U}_{[0,1]}$) möglichst nahe kommt.

Pseudozufall: Mersenne Twister

Moderner Zufallszahlengenerator, der auch in Matlab zur Erzeugung gleichverteilter Zufallszahlen Verwendung findet.
Es sei $\mathbb{F}_2$ der Körper der Charakteristik 2, also $\mathbb{F}_2 = \{0, 1\}$ mit der Addition $\oplus$ und der Multiplikation $\odot$.

Vorgegeben seien Parameter $\omega \in \mathbb{N}, m, n \in \mathbb{N}$, $1 \leq m < n$ und $r \in \{0, \ldots, \omega - 1\}$.

Notation:

$$\begin{aligned} y^l &= (y_1, \ldots, y_r, 0, \ldots, 0), \\ z^u &= (0, \ldots, 0, z_{r+1}, \ldots, z_\omega), \\ (y^l | z^u) &= (y_1, \ldots, y_r, z_{r+1}, \ldots, z_\omega). \end{aligned}$$

Pseudozufall: Mersenne Twister

Notation:

$$\begin{aligned} y^l &= (y_1, \dots, y_r, 0, \dots, 0), \\ z^u &= (0, \dots, 0, z_{r+1}, \dots, z_\omega), \\ (y^l | z^u) &= (y_1, \dots, y_r, z_{r+1}, \dots, z_\omega). \end{aligned}$$

Mit $\omega \times \omega$ Matrix A und $(n-1)$ Startwerten $x_0, \dots, x_{n-1} \in \mathbb{F}_2^\omega$ ist der Mersenne-Twister der folgende rekursive Algorithmus:

$$x_{k+n} = x_{k+m} \ \oplus^\omega \ \left(x_k^l | x_{k+1}^u \right) \odot^\omega A, k \in \mathbb{N}_0.$$

Dabei Addition und (Matrix)Multiplikation mit $\oplus^\omega$ und $\odot^\omega$ in $\mathbb{F}_2^\omega$.

Für $(\omega, n, m, r) = (32, 624, 397, 31) \rightsquigarrow$ Periodenlänge $2^{19937} - 1$.

Quasi–Zufallszahlen

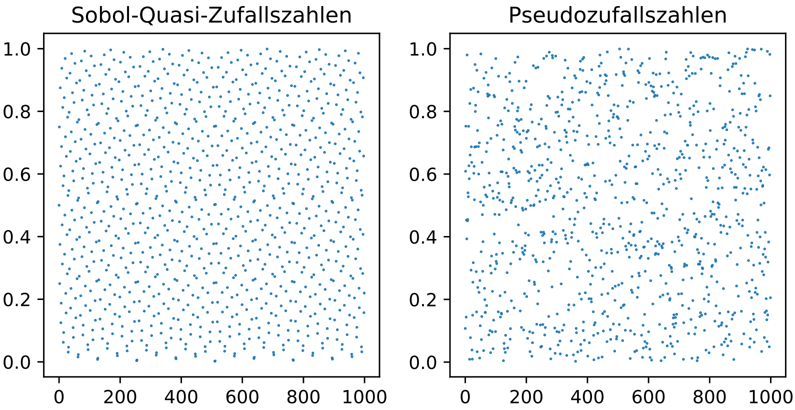
Ziel bei der Generierung einer Folge von Quasi-Zufallszahlen $x_1, \ldots, x_N$ ist die Minimierung der Diskrepanz

$$D_N(x_1, \ldots, x_N) = \sup_{u \in [0,1]} \left| \frac{|\{x_i : i = 1, \ldots, N\,,\ x_i \in [0, u)\}|}{N} - u \right| .$$

Quasi-Zufallszahlen (Halton-, Sobol-Folgen): $D_N \leq C(\log N/N)$

Für Pseudozufallszahlen indes $D_N \simeq N^{-1/2}$ nach ZGWS

$\Rightarrow$ kleinere Fehler bei Quasi-Monte Carlo Integration


Sobol-Quasi-Zufallszahlen
Pseudozufallszahlen
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000

Quasi-Zufall: Halton-Folge

Man wähle eine Primzahl p als Basis und einen Startwert $m \neq 0$ und stelle m zur Basis p dar:

$$m = \sum_{j=0}^{k} a_j p^j \ .$$

Die Haltonzahlen sind dann gegeben durch

$$h = \sum_{j=0}^{k} a_j p^{-j-1}.$$

Man verfährt analog mit $(m+1), \ldots$.

Quasi-Zufall: Halton-Folge

Die folgende Tabelle enthält die ersten drei Haltonzahlen zum Startwert $m = 3$ für $p = 2$.

m	binäre Darstellung	h
3	11	$3/4$
4	100	$1/8$
5	101	$5/8$
$\vdots$	$\vdots$	$\vdots$

Erzeugung Bernoulli–verteilter Zufallszahlen

Gegeben eine Zufallsvariable $U \sim \mathcal{U}_{[0,1]}$ folgt

$$X := T(U) = \begin{cases} 1 & \text{für} \quad U \leq p \\ 0 & \text{für} \quad U > p \end{cases}$$

ist Bernoulli-verteilt mit Parameter $p \in (0, 1)$.

Erzeugt man iid. $U_1, \ldots, U_n \sim \mathcal{U}_{[0,1]}$ und addiert die resultierenden X_i, so ist die Summe binomialverteilt mit Parametern n und p.

Ähnliche Diskretisierungsmethoden sind für andere diskrete Verteilungen anwendbar!

Die Inversionsmethode für univariate Verteilungen

Definition

Sei $F : \mathbb{R} \to [0,1]$ eine Verteilungsfunktion. Die Funktion

$$F^{-1}(u) := \begin{cases} \inf\{x \mid F(x) \geq u\}, & u \in (0,1] \\ \sup\{x \in \mathbb{R} \mid F(x) = 0\}, & u = 0 \end{cases}$$

heißt Quantilfunktion oder Verallgemeinerte Inverse von F.

Die Inversionsmethode für univariate Verteilungen

Lemma

Mit $U \sim \mathcal{U}_{[0,1]} \Rightarrow X = F^{-1}(U) \sim F$ für jede Verteilungsfunktion F. D.h. die Zufallsvariable $F^{-1}(U)$ ist verteilt nach der zu F gehörenden Verteilung.

Beweis:

$$\mathbb{P}[X \leq x] = \mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[U \leq F(x)] = F(x).$$

Die Inversionsmethode für univariate Verteilungen

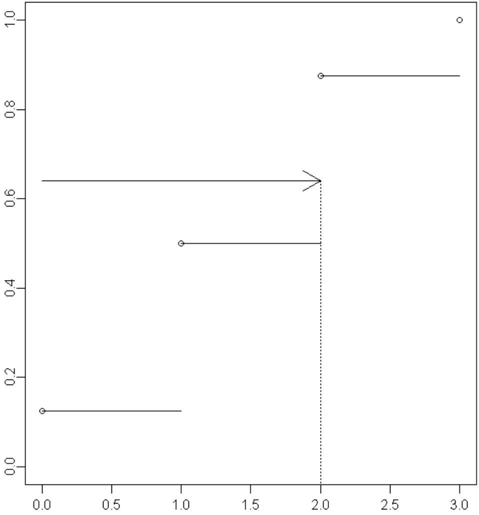
Seien $p_i, i \in J$ die Gewichte einer diskreten Verteilung zu den Werten $m_i, i \in J$, wobei $m_1 < m_2 < \ldots < m_{|J|}$. Die rechtsstetige Treppenfunktion

$$F(x) = \sum_{i: m_i \leq x} p_i$$

ist die zugehörige Verteilungsfunktion. Wenn U eine auf dem Einheitsintervall gleichverteilte Zufallszahl ist, wird durch

$$X := \min\{t,\ U \leq F(t)\}$$

eine Zufallszahl der diskreten Verteilung F erzeugt.



0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Erzeugung multivariat normalverteilter Zufallszahlen

Eine reellwertige Zufallsvariable heißt $\mathcal{N}(\mu, \sigma^2)$-verteilt, wenn sie die stetige Dichte

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/(2\sigma^2))$$

bezüglich des Lebesguemaßes besitzt.
Ist Z standardnormalverteilt, so ist

$$X = \mu + \sigma Z$$

normalverteilt mit Parametern μ und σ^2.

Erzeugung multivariat normalverteilter Zufallszahlen

Box-Muller-Methode: $U, V \overset{i.i.d.}{\sim} \mathcal{U}_{[0,1]} \Rightarrow$

$$(X, Y) := \sqrt{-2 \log U}\big(\cos 2\pi V\,,\ \sin 2\pi V\big) \quad \overset{i.i.d.}{\sim}\ \mathcal{N}(0,1).$$

Polarkoordinaten: $R^2 := X^2 + Y^2 \sim \mathrm{Exp}(1/2)$ und
$\theta := \arctan(Y, X) = \arg(X + iY) \sim \mathcal{U}_{[0,2\pi]}$ sind unabhängig!

Marsaglia-Bray Variante für BM: Sind (U, V) uniform auf dem Einheitskreis verteilt (mit Verwerfen erzeugen), so folgt

$$(X, Y) := \sqrt{\frac{-2 \log(U^2 + V^2)}{U^2 + V^2}}(U, V) \quad \overset{i.i.d.}{\sim}\ \mathcal{N}(0,1).$$

Erzeugung multivariat normalverteilter Zufallszahlen

Definition

X ist multivariat $\mathcal{N}(\mu, Q)$-verteilt in $\mathbb{R}^d$ mit Mittelwertvektor m und Kovarianzmatrix Q, falls für alle $\theta \in \mathbb{R}^d$ gilt

$$\theta^\top X \sim \mathcal{N}(\theta^\top m, \theta^\top Q\,\theta)\text{-univariat normalverteilt.}$$

Falls Matrix Q diagonal ist, werden Koordinaten von X unabhängig simuliert. Und anderenfalls?

Erzeugung multivariat normalverteilter Zufallszahlen

Definition

X ist multivariat $\mathcal{N}(\mu, Q)$-verteilt in $\mathbb{R}^d$ mit Mittelwertvektor m und Kovarianzmatrix Q, falls für alle $\theta \in \mathbb{R}^d$ gilt

$$\theta^\top X \sim \mathcal{N}(\theta^\top m, \theta^\top Q\, \theta)\text{-univariat normalverteilt.}$$

Für $Z \sim \mathcal{N}(0, I_d)$ standart multivariat normal ist

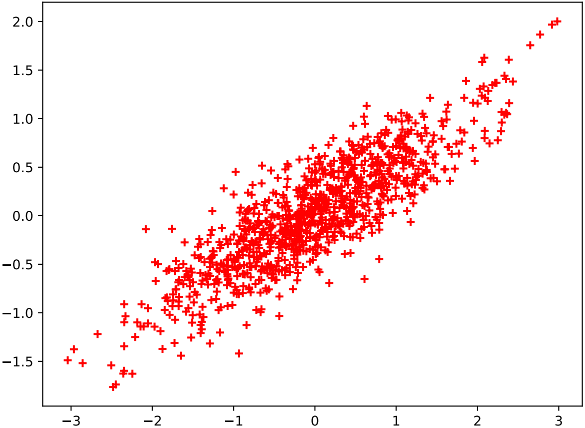
$$X := m + BZ \sim \mathcal{N}(m, BB^\top) \quad m \in \mathbb{R}^d\,,\ B \in \mathbb{R}^{d \times d}\,.$$

Für ggb. Kovarianzmatrix Q liefert z.B. Cholesky-Zerlegung $Q = ADA^\top = BB^\top$ (oder $B = AD^{1/2}$) eine Matrix B s.d. $X := m + BZ \sim \mathcal{N}(m, Q)$-normalverteilt ist.

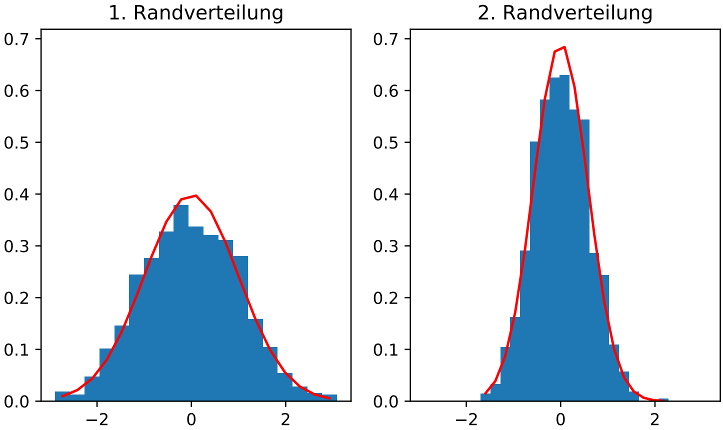
Erzeugung multivariat normalverteilter Zufallszahlen

Scatterplot einer Stichprobe von 1000 Pseudozufallszahlen einer 2-dimensionalen $\mathcal{N}(0, Q)$-Normalverteilung zu Kovarianz

$$Q = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.333 \end{pmatrix}$$



Dichte-Histogramme der empirischen Randverteilungen verglichen mit theoretischen Randverteilungsdichten:



Übersicht

Monte Carlo-Integration

Näherungsweise Berechnung von

$$\begin{aligned}\int_a^b g(x)\,\mathrm{d}x &= \int_a^b h(x)f(x)\,\mathrm{d}x \\ &= \int h(x)\mathbb{P}_f[\,\mathrm{d}x] = \mathbb{E}_f[h(X)]\end{aligned}$$

durch den Mittelwert einer endlichen Folge von i.i.d. erzeugten Zufallszahlen X_i mit Verteilung $\mathbb{P}_f$:

$$S := \frac{1}{n}\sum_{i=1}^{n} h(X_i)\,.$$

Crude Monte Carlo Integration

Speziell mit $X_i \overset{i.i.d.}{\sim} \mathcal{U}_{[a,b]}$ ergibt sich der Schätzer

$$S = \frac{1}{n} \sum_{i=1}^{n} g(X_i)(b-a)\,.$$

Der Schätzer ist erwartungstreu, da

$$\mathbb{E}[S] = \frac{b-a}{n} \sum_{1}^{n} \mathbb{E}[g(X_i)] = \int_a^b g(x)\,\mathrm{d}x$$

und nach dem Gesetz der großen Zahlen konsistent mit Varianz

$$\frac{(b-a)^2}{n^2} \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(g(X_i)) = \frac{b-a}{n} \int_a^b \Big(g(x) - \int_a^b g(t)\,\mathrm{d}t \Big)^2 \mathrm{d}x.$$

Monte-Carlo-Bänder

Mit MC können empirisch durch wiederholte Simulation von (funktionalen) Statistiken für diese sogenannte Monte-Carlo-Bänder erzeugt werden, um die Variabilität der Statistik und Konfidenzbereiche zu veranschaulichen.

Beispiel Monte-Carlo-Bänder für eine empirische Verteilungsfunktion einer (Pseudo) $U[0,2]$-iid. Stichprobe

Monte-Carlo-Bänder

Beispiel Monte-Carlo-Bänder für eine empirische Verteilungsfunktion einer (Pseudo) $U[0,2]$-iid. Stichprobe