BZQ II: Stochastikpraktikum

Block 1: Monte-Carlo-Methoden, Zufallszahlen, Statistische Tests

Randolf Altmeyer

November 22, 2016

Überblick

1. Monte-Carlo-Methoden, Zufallszahlen, statistische Tests
2. Lineares Modell, Klassifikation
3. Nichtparametrische Verfahren
4. Markov-Chain-Monte-Carlo-Verfahren
5. Simulation stochastischer Prozesse

Vielleicht: PCA, mehr Datenanalyse

Literatur

- Dirk Kroese, T. Taimre, Z.I. Botev: *Handbook of Monte Carlo Methods*
- Christian Robert, George Casella: *Introducing Monte Carlo Methods with R*
- Christian Robert, George Casella: *Monte Carlo Statistical Methods*
- Trevor Hastie, Robert Tibshirani, Jerome H. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*
- Mathias Trabs, Markus Reiß, Moritz Jirak: *Methoden der Statistik* (Skript)

Die Monte-Carlo-Methode

- Zufallsvariable X mit Verteilung P^X
- **Ziel**: für Funktion g werte Erwartungswerte der Form

$$\mathbb{E}\left[g\left(X\right)\right] = \int g\left(x\right) dP^X\left(x\right)$$

  aus
- Ursprung des Namens: Johann von Neumann, Stanislaw Ulam (1946), Codename für ein Geheimprojekt im "Los Alamos Scientific Laboratory" in Anlehnung an das Kasino "Monte Carlo" in Monaco

Die Monte-Carlo-Methode

Idee:

- erzeuge iid Zufallszahlen $(X_k)_{k\geq 1}$ mit $X_k \overset{d}{\sim} P^X$
- definiere $S_n := \frac{1}{n}\sum_{k=1}^{n} g(X_k)$
- starkes Gesetz der großen Zahlen (wenn $g(X) \in L^1$):

$$S_n \xrightarrow{f.s.} \mathbb{E}[g(X)],$$

  d.h. $S_n \approx \mathbb{E}[g(X)]$
- S_n ist *erwartungstreu* (d.h. $\mathbb{E}[S_n] = \mathbb{E}[g(X)]$ und *konsistent*
- Schätzabweichung: wenn $X \in L^2$, dann gilt für $\varepsilon > 0$

$$\mathbb{P}\left(|S_n - \mathbb{E}[g(X)]| > \frac{\varepsilon}{\sqrt{n}}\right) \leq \frac{\frac{1}{n^2}\sum_{k=1}^{n} Var(g(X))}{\frac{\varepsilon^2}{n}} = \frac{Var(g(X))}{\varepsilon^2}$$

Monte-Carlo Integration

- Gegeben: Funktion g
- **Ziel**: berechne $\int_0^1 g(x)\,dx$
- **Idee**:
  - es gilt $\mathbb{E}[g(U)] = \int_0^1 g(x)\,dx$ für $U \overset{d}{\sim} U([0,1])$
  - werte $\mathbb{E}[g(U)]$ mit Monte-Carlo-Simulationen aus
  - Vorteile: maximale Konvergenzrate $n^{-1/2}$ ist unabhängig von g und von der Dimension (anders als bei anderen Quadraturformeln)
  - Nachteil: nutzt eventuelle Glattheit von g nicht aus

Fehlerauswertung für die Monte-Carlo-Methode

- mit zentralem Grenzwertsatz:
  - $X_k \overset{iid}{\sim} \mathbb{P}^X$ (iid = independent, identically distributed), $S_n := \frac{1}{n}\sum_{k=1}^{n} g(X_k)$, $\mu = \mathbb{E}[S_n] = \mathbb{E}[g(X_1)]$, $\sigma_n = \mathrm{Var}^{1/2}(S_n)$
  - $\sqrt{n}\left(\frac{S_n-\mu}{\sigma_n}\right) \xrightarrow{d} N(0,1) \Rightarrow \mathbb{P}\left(\sqrt{n}\left|\frac{S_n-\mu}{\sigma_n}\right| \geq 2\right) \approx 0.05$
  - verwende daher als Konfidenzband: $\mu - 2\frac{\sigma_n}{\sqrt{n}} \leq S_n \leq \mu + 2\frac{\sigma_n}{\sqrt{n}}$
  - in Abhängigkeit von n:
    - $\mu \approx S_n$
    - $\sigma_n^2 = \frac{1}{n^2}\sum_{k=1}^{n} \mathrm{Var}g(X_k) = \frac{1}{n}\mathrm{Var}g(X_1) \approx \frac{1}{n}\sum_{k=1}^{n}(X_k - S_n)^2$
- durch Simulation der Varianz:
  - führe m verschiedene Monte-Carlo-Simulationen durch und erhalte $S_n^{(1)}, \ldots, S_n^{(m)}$
  - $\mu \approx \frac{1}{n}\sum_{k=1}^{m} S_n^{(k)}$, $\sigma_n^2 \approx \frac{1}{m}\sum_{k=1}^{m}\left(S_n^{(k)} - \mu\right)^2$

Erzeugen von Zufallszahlen

- **Ziel**:
  - gegeben: Verteilung $\mathbb{P}^X$, z.B. über Verteilungsfunktion F^X oder Dichte f^X
  - erzeuge Zahlen $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}^X$ (iid = independent, identically distributed)
- mögliche Methoden:
  - physikalische Methoden (Hintergrundstrahlung, Atomzerfall)
  - am Computer: mit **Zufallszahlengenerator** (= *deterministischer* Algorithmus, der pseudo-zufällige Zahlen erzeugt)
- Eigenschaften eines "guten" Zufallszahlengenerators:
  - besteht viele statistische Tests (auf Unabhängigkeit, auf Verteilungsannahme, etc., siehe auch Marsaglia's *Die hard tests*)
  - reproduzierbar (ohne alle Zahlen zu speichern)
  - schnell, "billig"
  - lange Periode

Methode 1: Zufallszahlengeneratoren für $U([0,1])$

Idee:

- Funktion $f : [0,1] \to [0,1]$
- "Seed" U_0
- erzeuge rekursiv $U_n = f(U_{n-1})$
- Computer ist *diskret*, d.h. es gibt nur endlich viele mögliche Werte
  $\Rightarrow$ Werte wiederholen sich nach einer bestimmten Zeit/Periode

Beispiele

1. "chaotische" dynamische Systeme
   - sollen stark von Störungen der Anfangsbedingungen abhängen
   - Beispiel: *logistische Funktion* $X_n = \alpha X_{n-1}(1 - X_{n-1})$,
     $U_n = \frac{2}{\pi} \sin^{-1}\left(\sqrt{X_n}\right)$
   - keine Garantie für gute Eigenschaften, hängen sehr stark von Rundungsfehlern ab (können sogar konvergieren)

Methode 1: Zufallszahlengeneratoren für $U([0,1])$

2. Lineare Kongruenzgeneratoren

- wähle $a, b, m \in \mathbb{N}$ fest
- Seed $X_0 \in \{0, \ldots, m-1\}$
- Regel: $X_n = aX_{n-1} + b \,(\mod m) \Rightarrow U_n = \frac{X_n}{m} \in [0,1]$
- Periode $\leq m$, Seed und b nicht so wichtig
- können extrem schlechte Eigenschaften haben je nach Wahl von a, m
  Beispiel: $a = 65539$, $m = 2^{31}$. Dann ist $a = 2^{16} + 3$ und damit (mod m gerechnet)

$$\begin{aligned} X_{n+2} &= \left(2^{16}+3\right) X_{n+1} = \left(2^{16}+3\right)^2 X_n \\ &= \left(2^{32} + 6 \cdot 2^{16} + 9\right) X_n = \left(6 \cdot 2^{16} + 9\right) X_n \\ &= \left(6 \cdot \left(2^{16}+3\right) - 9\right) X_n \\ &= 6X_{n+1} - 9X_n \end{aligned}$$

Aus dem *Satz von Marsaglia* folgt, dass die Tripel (X_n, X_{n+1}, X_{n+2}) stets auf einer von 15 verschiedenen Hyperebenen liegen.

Methode 1: Zufallszahlengeneratoren für $U([0,1])$

3. Mersenne-Twister
   - zuverlässiger Zufallsgenerator, besteht viele statistische Tests,
   - Periode $2^{19937} - 1$
4. Kombination von Zufallszahlengeneratoren
   - wesentlich größere Perioden
   - kombinieren die guten Eigenschaften von Generatoren
   - Beispiel: $U_n = \frac{X_n}{m_1} + \frac{Y_n}{m_2} (\mathrm{mod}\, 1)$ für zwei lineare Kongruenzgeneratoren X_n, Y_n

- Übersicht über bekannte Zufallszahlengeneratoren und ihre Eigenschaften: http://random.mat.sbg.ac.at/results/karl/server/
- ab jetzt nehmen wir an, dass wir schnell, billig und zuverlässig Folgen von unabhängigen $Unif([0,1])$-verteilten Zufallsvariablen erzeugen können

Methode 2: Inversionsmethode

- **Stochastik 1**: Sei X reellwertige Zufallsvariable mit Verteilungsfunktion F.
  - definiere $F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$
  - es gilt: $F(F^{-1}(u)) \geq u$ und $F^{-1}(F(x)) \leq x$, d.h. $\{(u,x) : F^{-1}(u) \leq x\} = \{(u,x) : F(x) \geq u\}$ und daher $\mathbb{P}\left(F^{-1}(U) \leq x\right) = \mathbb{P}(U \leq F(x)) = F(x)$.
  - $U = F(X) \overset{d}{\sim} \mathit{Unif}([0,1])$
  - wenn $V \overset{d}{\sim} \mathit{Unif}([0,1])$, dann ist $F^{-1}(V) \overset{d}{\sim} X$
- es genügt also (theoretisch) uniforme Zufallsvariablen zu erzeugen $\Rightarrow$ das implizite Tripel $(\Omega, \mathcal{F}, \mathbb{P})$ kann stets realisiert werden über $\left([0,1], \mathcal{B}_{[0,1]}, \mathit{Unif}([0,1])\right)$
- **aber**: i.A. ist F^{-1} nicht explizit verfügbar
- **Beispiel**: Sei $X \overset{d}{\sim} \mathit{Exp}(1)$. Dann ist $F(x) = 1 - e^{-x}$, d.h. $F^{-1}(u) = -\log(1-u)$. Wenn $U \overset{d}{\sim} \mathit{Unif}([0,1])$, dann ist $-\log U \sim \mathit{Exp}(1)$ $(1 - U \sim \mathit{Unif}([0,1]))$.

Methode 3: Spezielle Methoden

1. **Erzeuge neue Zufallsvariablen aus alten:**
   - $U \overset{d}{\sim} U([0,1]) \Rightarrow aU + b \overset{d}{\sim} U([b, a+b])$
   - $X_1, \ldots, X_k \overset{d}{\sim} Exp(1) \Rightarrow \sum_{i=1}^{k} X_i \overset{d}{\sim} Gamma(k,1)$ **(Plot)**
   - $X_1, \ldots, X_k \overset{d}{\sim} N(0,1) \Rightarrow \sum_{i=1}^{k} X_i^2 \overset{d}{\sim} \chi^2(k)$
   - Box-Muller-Methode: $U_1, U_2 \overset{d}{\sim} U([0,1])$
     $\Rightarrow (X_1, X_2) = (\sqrt{-2\log(U_1)}\cos(2\pi U_2), \sqrt{-2\log(U_1)}\sin(2\pi U_2)) \overset{d}{\sim} N(0, I_2)$
   - $X \overset{d}{\sim} N(0, \Sigma)$, $\Sigma \in \mathbb{R}^{p \times p}$,$\mu \in \mathbb{R}^p$, $A \in \mathbb{R}^{q \times p} \Rightarrow \mu + AX \overset{d}{\sim} N(\mu, A^\top \Sigma A)$

Methode 3: Spezielle Methoden

2. **Diskrete Verteilungen (hier: $X \in \mathbb{N}_0$)**
   - Idee: Verteilungsfunktion ist in diesem Fall immer invertierbar
   - berechne (und speichere) $p_k = F(k) = \mathbb{P}(X \leq k)$
   - erzeuge $U \overset{d}{\sim} Unif([0,1])$ und setze $X = k$, wenn $p_{k-1} < U < p_k$
   - es ist ineffizient immer vorne bei $k = 0$ mit dem Testen anzufangen, vorallem wenn X viele Werte annimmt
     Beispiel: $X \overset{d}{\sim} \text{Poiss}(\lambda)$ mit $\lambda = 100 \Rightarrow$ Werte fast ausschließlich zwischen $\lambda \pm 3\sqrt{\lambda} = [70, 130]$
     mögliche Lösung: ignoriere alles außerhalb bestimmter Bereiche mit kleiner Wahrscheinlichkeit

Methode 4: Accept-Reject-Sampling

- Motivation für allgemeinere Methoden (Caselle)
- **Ziel:** erzeuge $X \overset{d}{\sim} \mathbb{P}_f$ mit Wahrscheinlichkeitsdichte f
- **Gegeben:** *Kandidatendichte* g, so dass man von $\mathbb{P}_g$ "leicht" Zufallszahlen erzeugen kann und so dass $\frac{f(x)}{g(x)} \leq M$ für alle $x \in \mathbb{R}$ und eine Konstante M
- **Accept-Reject (Verwerfungsmethode):**
  - *Schritt 1*: erzeuge $U \overset{d}{\sim} Unif([0,1])$, $Y \overset{d}{\sim} P_g$, $U \perp Y$ (d.h. U und Y unabhängig)
  - *Schritt 2*:
    - wenn $U \leq f(Y)/Mg(Y)$, dann *akzeptiere* $X := Y$
    - andernfalls *lehne* Y *ab* und kehre zu Schritt 1 zurück

Methode 4: Accept-Reject-Sampling

- **Warum funktioniert das?**

$$\mathbb{P}(X \leq x) = \mathbb{P}_g\left(Y \leq x | U \leq \frac{f(Y)}{Mg(Y)}\right) = \frac{\mathbb{P}_g\left(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}\right)}{\mathbb{P}_g\left(U \leq \frac{f(Y)}{Mg(Y)}\right)} =$$

$$\frac{\int_{-\infty}^{x}\int_{0}^{f(y)/Mg(y)} dug(y)dy}{\int_{-\infty}^{\infty}\int_{0}^{f(y)/Mg(y)} dug(y)dy} = \frac{\frac{1}{M}\int_{-\infty}^{x} f(y)dy}{\frac{1}{M}\int_{-\infty}^{\infty} f(y)dy} = \mathbb{P}_f((-\infty, x])$$

- unabhängig von der Dimension, f und g müssen nur bis auf Konstanten bekannt sein, die Konstante M muss nicht "scharf" sein
- größtes Problem: finden von g (intuitiv wird es schwieriger von g zu sampeln, wenn $M \to 1$)
- **Akzeptanzwahrscheinlichkeit:**
  - $\mathbb{P}_g\left(U \leq \frac{f(Y)}{Mg(Y)}\right) = \int_{-\infty}^{\infty}\int_{0}^{f(y)/Mg(y)} dug(y)dy = \frac{1}{M}\int_{-\infty}^{\infty} f(y)dy = \frac{1}{M}$
  - je kleiner M, desto höher ist die Akzeptanzwahrscheinlichkeit, d.h. desto weniger Samples werden abgelehnt

Methode 4: Accept-Reject-Sampling

- **Beispiel**:
  - erzeuge $X \overset{d}{\sim} Beta(4,3)$, d.h.
    $f(x) = \frac{1}{B(4,3)} x^{4-1}(1-x)^{3-1} = 60x^3(1-x)^2$ mit $B(4,3) = 1/60$.
  - Inversionsmethode ist nicht möglich analytisch
  - für Accept-Reject-Methode wähle $g(y) = 1$, d.h. $Y \overset{d}{\sim} Unif([0,1])$
  - die Konstante M ergibt sich aus

$$\begin{array}{rcl} f(x) & \leq & Mg(x) = M \\ \Leftrightarrow \quad 60x^3(1-x)^2 & \leq & M \end{array}$$

  für $x \in [0,1]$, d.h. $M \approx 2.2$.

Überblick

1. Zufallszahlengeneratoren für $U([0,1])$
2. Inversionsmethode
3. Spezielle Methoden
4. Accept-Reject-Sampling
5. später: Importance-Sampling, Markov-Chain-Monte-Carlo

(Naive) Einführung in Statistische Tests

- Grundidee: wir wollen gegeben die Zufallszahlen $X_1, \ldots, X_n$ entscheiden, ob diese bestimmte Eigenschaften haben
- Eigenschaften werden über die Verteilung $X_1, \ldots, X_n \overset{d}{\sim} P_\theta$ bestimmt, wobei $\theta \in \Theta$ (= Parametermenge)
- Beispiel:
  - beobachte Zufallszahlen $X_1, \ldots, X_n \overset{d}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$. Was ist das wahre θ?
  - beobachte Zufallszahlen $X_1, \ldots, X_n \overset{d}{\sim} \mathbb{P}_F$ mit Verteilungsfunktion F. Was ist das wahre F?
- Nullhypothese H_0 vs. Alternative H_1:
  - zerlege Θ in $H_0 \uplus H_1$ und *entscheide,* ob $\theta \in H_0$ oder $\theta \in H_1$

(Naive) Einführung in Statistische Tests

- Statistischer Test:
  - Zufallsvariable $\varphi_n : \Omega \to [0,1]$ heißt Test von H_0 gegen H_1, wenn $\varphi_n(X_1, \ldots, X_n) = 1$ bedeutet, dass H_0 verworfen wird (d.h. $\theta \in H_1$) und $\varphi_n(X_1, \ldots, X_n) = 0$ bedeutet, dass H_0 nicht verworfen wird (d.h. $\theta \in H_0$).
  - wähle φ_n so, dass $\mathbb{E}_\theta(\varphi_n(X_1, \ldots, X_n)) \leq \alpha$ für kleines $\alpha > 0$ und $\theta \in H_0$ (d.h. φ_n hat Signifikanzniveau/Level α) und $\mathbb{E}_\theta(\varphi_n(X_1, \ldots, X_n))$ ist maximal, wenn $\theta \in H_1$ (Power von φ_n)
- oft: $\varphi_n(X_1, \ldots, X_n) = \mathbf{1}(T(X_1, \ldots, X_n) \geq c_\alpha)$ für eine Zufallsvariable $T(X_1, \ldots, X_n)$ und "kritische Werte" c_α
- wir lehnen ab, wenn $T(X_1, \ldots, X_n)$ zu groß, d.h. unter der Nullhypothese unwahrscheinlich wird
- wenn $T(X_1, \ldots, X_n)$ unter der Nullhypothese eine bekannte Verteilung hat, dann kann man c_α über Quantile bestimmen
- Beispiel:
  - $X_1, \ldots, X_n \overset{d}{\sim} N(\theta, 1)$, $H_0 = \{0\}$, $H_1 = \mathbb{R} \backslash \{0\}$
  - $T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{k=1}^{n} X_k$
  - $c_\alpha = 1 - \alpha$-Quantil von $N(0,1)$

Statistischer Test für Zufallszahlen

Kolmogorov-Smirnov-Test (Verteilungstest)

- betrachte empirische Verteilungsfunktion $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i)$ mit "wahrer" Verteilungsfunktion F^X
- Stochastik 1: $D_n = \sup_x |F_n(x) - F^X(x)| \to 0$ f.s. für $n \to \infty$ (Satz von Glivenko-Cantelli)
- Stochastik 2: $\sqrt{n} D_n \xrightarrow{d} K$, wobei K die Kolmogorov-Verteilung hat
- für den Test:
  - $\Theta =$ Menge aller Verteilungsfunktionen
  - $H_0 = \{F^X\}$, $H_1 = \Theta \backslash \{F^X\}$
  - $T(X_1, \ldots, X_n) = D_n$
  - bestimme c_α durch $\mathbb{P}_{F^X}\left(\sqrt{n} D_n \geq c_\alpha\right) = \alpha$
- Durchführen des Tests mit "konkreten" Zufallszahlen:
  - Auswerten von D_n mit den Daten
  - wenn $\sqrt{n} D_n \geq c_\alpha$, dann lehne ab, sonst lehne nicht ab

Statistischer Test für Zufallszahlen

- in *R*:
  - Testaufruf mit dem Kommando ks.test(x,y), wobei z.B. $y =$**pnorm**
  - Rückgabe eines *p*-Wertes. Er heißt auch *beobachtetes Signifikanzniveau* und erfüllt $p \equiv p(X_1, \ldots, X_n) = \inf\{\alpha : T(X_1, \ldots, X_n) \geq c_\alpha\}$. Er ist zufällig (hängt von den Daten ab) und ist das kleinste Signifikanzniveau bei dem wir noch ablehnen würden für diese Daten. Er hängt damit implizit mit den c_n zusammen.
  - kurz gesagt: wenn $p < \alpha$, dann ist das Ergebnis *signifikant* und wir lehnen ab