Stochastik–Praktikum
Markov Chain Monte Carlo–Methoden

Thorsten Dickhaus

Humboldt-Universität zu Berlin

14.10.2010

Problemstellung

Wie kann eine Zufallsstichprobe am Computer simuliert werden, deren Verteilung aus einem komplizierten Bildungsgesetz (z. B. hierarchisch Bayes) herrührt?

Definition

Eine Markov Chain Monte Carlo (MCMC)–Methode zur Simulation einer Wahrscheinlichkeitsverteilung mit Dichte f produziert eine ergodische Markov–Kette, welche die Verteilung mit Dichte f als stationäre Verteilung aufweist.

Übersicht

1. Grundlegendes über Markov–Ketten
2. Der Metropolis–Hastings–Algorithmus
3. Anwendung auf das Logit–Modell
4. Der Gibbs–Sampler

Im Folgenden bezeichnet $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ einen messbaren Raum.

Definition

Ein Übergangskern ist eine Abbildung $\mathcal{K} : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}$,
für die gilt:

- $\mathcal{K}(x, \cdot\,)$ ist ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathcal{X}) \;\; \forall x \in \mathcal{X}$
- $\mathcal{K}(\,\cdot\,, B)$ ist messbar $\forall B \in \mathcal{B}(\mathcal{X})$

Definition

Ein stochastischer Prozess in diskreter Zeit $X_0, \ldots, X_n, \ldots$ heißt Markov–Kette (MC) (X_n) auf $\mathcal{X}$, falls $\forall B \in \mathcal{B}(\mathcal{X})$

$$P\left(X_{k+1} \in B | x_0, \ldots, x_k\right) = P\left(X_{k+1} \in B | x_k\right) = \int_B \mathcal{K}(x_k, dx)$$

für $k \in \mathbb{N}$ mit einem Übergangskern $\mathcal{K}$ gilt.

Definition (Sei $\mathcal{X}$ abzählbar.)

a) Ersteintrittszeit in $x \in \mathcal{X}$:
$T^x := \min\{n > 0 | X_n = x\}$, $f(x,y) := P_x(T^y < \infty)$ $x, y \in \mathcal{X}$

b) (X_n) heißt zeithomogen, falls $\forall k \in \mathbb{N}$:

$$P(X_{k+1} \in B | x_k) = P(X_1 \in B | x_0).$$

c) Ein Zustand $x \in \mathcal{X}$ heißt rekurrent, falls $f(x,x) = 1$.
Eine Markov–Kette (X_n) heißt rekurrent, falls jeder ihrer möglichen Zustände rekurrent ist. Ansonsten heißt (X_n) transient.

d) Eine Markov–Kette heißt irreduzibel, falls jeder Zustand y von jedem Zustand x aus erreichbar ist.

Satz

1. Falls MC (X_n) irreduzibel:
   Ein Zustand rekurrent $\Leftrightarrow$ (X_n) rekurrent
2. Ist eine Markov–Kette irreduzibel und rekurrent, dann existiert ein eindeutig bestimmtes invariantes Maß $\mu \equiv \mu\mathcal{K}$.

   Ist μ endlich, heißt die Markov–Kette positiv–rekurrent.

Korollar

MC irreduzibel und positiv–rekurrent, dann gilt:

$$\mathbb{E}[T^x] < \infty \quad \text{und} \quad \mu(x) = \left(\mathbb{E}[T^x]\right)^{-1} .$$

Ergodizität

Definition

Periode: $d_x := ggt\{n | \mathcal{K}^n(x,x) > 0\}$
MC heißt aperiodisch, wenn $d_x = 1 \ \forall x \in \mathcal{X}$.

Satz (Konvergenz ins Gleichgewicht)

Ist (X_n) eine irreduzible, positiv–rekurrente, aperiodische (ergodische) Markov–Kette mit stationärer Verteilung μ, so gilt für jede beliebige Startverteilung μ_0 von X_0:

$$\mu_0 \mathcal{K}^n \to \mu \ \text{ für } n \to \infty \ .$$

Markov Chain Monte Carlo–Algorithmen

Prinzip von MCMC–Algorithmen:

Für beliebige Startverteilungen wird eine ergodische Markov–Kette generiert.

Dabei wird der Übergangskern so gewählt, dass Konvergenz gegen eine festgelegte invariante Verteilung (die Zielverteilung) stattfindet.

Anwendungsbeispiel:

Näherungsweise Berechnung von $\int h(x)f(x)dx$ durch $T^{-1}\sum_{i=n_0}^{n_0+T} h(X_i)$, wobei f die stationäre Dichte ist und die sogenante „burn–in–Phase“ $\{1, \ldots, n_0 - 1\}$ ausgenommen wird.

Übersicht

Aufgabe: Simulation einer Stichprobe, die der Zielverteilung mit Dichte f folgt.

Gegeben sei $X^{(t)} = x^{(t)}$.

1. Generiere $Y_t \sim q(y|x^{(t)})$ gemäß einer proposal distribution q.
2. Acceptance–Rejection Schritt:

$$X^{(t+1)} = \begin{cases} Y_t & \text{mit Wahrscheinlichkeit } \rho(x^{(t)}, Y_t) \\ x^{(t)} & \text{mit Wahrscheinlichkeit } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

wobei

$$\rho(x, y) = \min\left(1, \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}\right).$$

Wahl der proposal distribution q

Metropolis–Hastings erzeugt eine irreduzible MC, falls

$$q(y|x) > 0 \ \forall (x, y) \in \operatorname{supp}(f) \times \operatorname{supp}(f).$$

Aperiodizität gilt, falls

$$\mathbb{P}\left(f(X^{(t)})q(Y_t|X^{(t)}) \leq f(Y_t)q(X^{(t)}|Y_t)\right) < 1.$$

Übergangskern des Metropolis-Hastings-Algorithmus'

Explizite Angabe von $\mathcal{K}$:

$$\mathcal{K}(x,y) = \rho(x,y)q(y|x) + \left(1 - \int \rho(x,y)q(y|x)dy\right)\delta_x(y).$$

Damit erfüllt (X_n) die detailed balance–Bedingung

$$\mathcal{K}(x,y)f(x) = \mathcal{K}(y,x)f(x) \quad \forall x,y.$$

Satz

Besitzt eine Markov–Kette die „detailed balance“ Eigenschaft, so ist sie irreduzibel und f Dichte der invarianten Verteilung.

Satz (Konvergenz von Metropolis–Hastings)

Ist die Metropolis–Hastings–Markov–Kette $(X^{(t)})$ f–irreduzibel, dann gilt

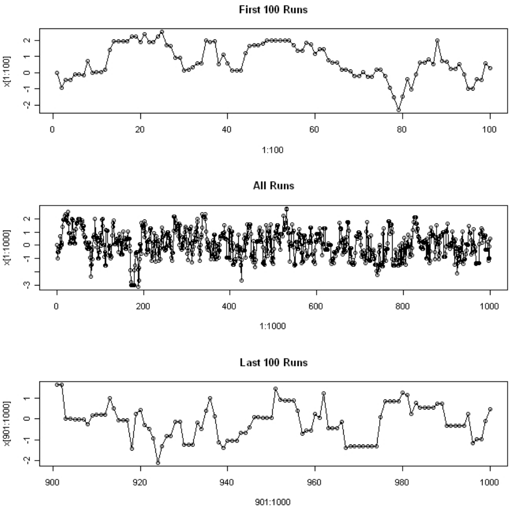
$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(X^{(t)}) = \int h(x)f(x)dx \;\; \forall h \in L^1(f),\; f - \text{ f. ü.}$$

Ist $(X^{(t)})$ aperiodisch, so gilt außerdem

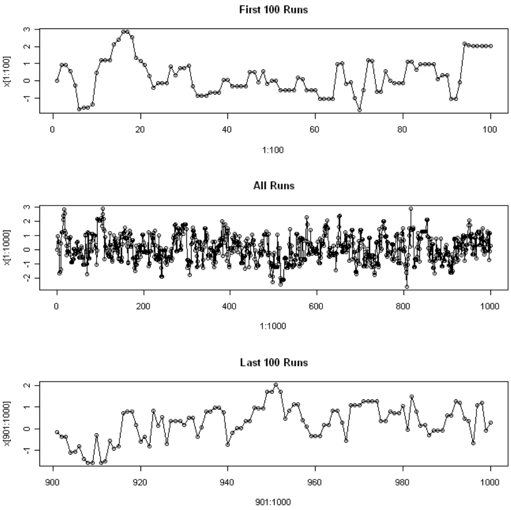
$$\lim_{n\to\infty} \left\| \int \mathcal{K}^n(x,\cdot)\mu(dx) - f \right\|_{TV} = 0$$

für jede Startverteilung μ.

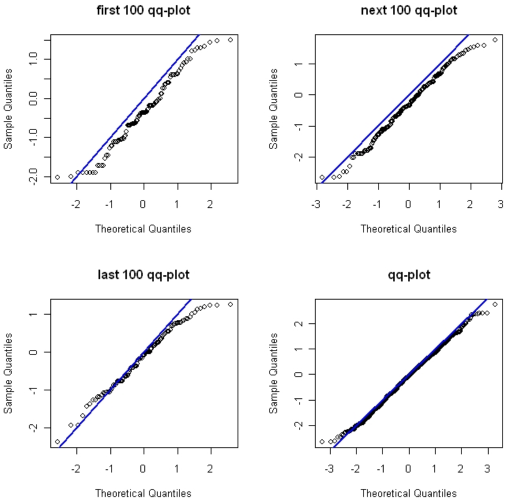
Beispiel: Normalverteilung



Beispiel: Normalverteilung



Beispiel: Normalverteilung



Übersicht

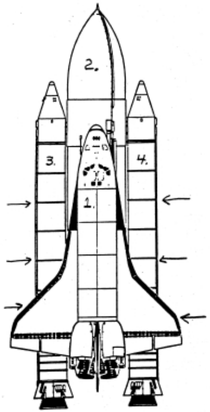
Challenger–Beispiel

Am 28. Januar 1986 explodierte die Raumfähre Challenger mit sieben Astronauten an Bord nach dem Start aufgrund von Materialermüdungserscheinungen an den O-Ringen.

Zum Startzeitpunkt herrschte eine ungewöhnlich niedrige Außentemperatur von 31° F (ca. 0° C).

Nachfolgend soll der Zusammenhang zwischen der Temperatur und der Überbeanspruchung der Dichtungsringe analysiert werden.

Challenger–Beispiel



X	Y	X	Y
66	0	67	0
70	1	53	1
69	0	67	0
68	0	75	0
67	0	70	0
72	0	81	0
73	0	76	0
70	0	79	0
57	1	75	0
63	1	76	0
70	1	58	1
78	0		

Logistische Regression

Modell: Logistische Regression (Logit–Modell)

Sei Y binäre, Bernoulli-verteilte Zufallsvariable mit Parameter $p(X)$ und X eine reellwertige Zufallsvariable:

$$\mathbb{P}(Y = 1) = p(X) = \frac{\exp(\alpha + X\beta)}{1 + \exp(\alpha + X\beta)} = 1 - \mathbb{P}(Y = 0).$$

Die Logit–Funktion $logit(p) = \log(p/(1 - p)) = \alpha + X\beta$ hängt also linear von X ab.

Logistische Regression

Bayes–Ansatz zur Schätzung der Parameter α und β:

$$L(\alpha, \beta) = \prod_{i=1}^{23} \mathbb{P}\left(Y_i = y_i | x_i\right)$$
$$= \prod_i \left(\frac{\exp\left(\alpha + \beta x_i\right)}{1 + \exp\left(\alpha + \beta x_i\right)} \right)^{y_i} \left(\frac{1}{1 + \exp\left(\alpha + \beta x_i\right)} \right)^{1-y_i}$$

mit A priori–Verteilung

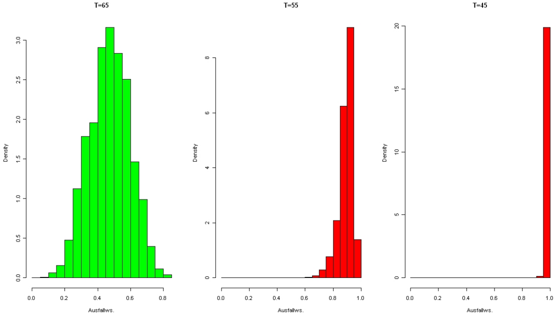
$$\pi(\alpha, \beta) = \pi_\alpha(\alpha|\beta)\pi_\beta(\beta).$$

(Hierarchisches Bayes-Verfahren)

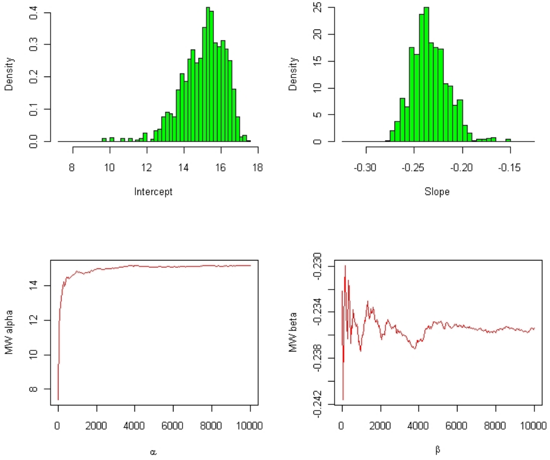
Metropolis–Hastings im Logit–Modell

Um von der Zielverteilung (A posteriori-Verteilung) zu samplen, wird nun ein Metropolis–Hastings–Algorithmus implementiert.

Resultierende Histogramme für die Ausfallwahrscheinlichkeit bei verschiedenen Temperaturen:



Konvergenz Metropolis–Hastings



Übersicht

Aufgabenstellung Gibbs-Sampler

$X = (X_1, \ldots, X_p)^T$ ein $\mathbb{R}^p$–wertiger Zufallsvektor.

Annahmen:

1. Weder von der gemeinsamen Verteilung mit Dichte f noch von den univariaten Randdichten $f_1, \ldots, f_p$ kann direkt simuliert werden.
2. Jedoch kann von den bedingten Verteilungen der

$$X_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \sim f_i(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

gesamplet werden.

Algorithmus: Gibbs-Sampler

Gegeben sei $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$. Generiere sukzessive

$$\begin{aligned}
X_1^{(t+1)} &\sim f_1(x_1 | x_2^{(t)}, \ldots, x_p^{(t)}),\\
X_2^{(t+1)} &\sim f_2(x_2 | x_1^{(t+1)}, x_2^{(t)}, \ldots, x_p^{(t)}),\\
&\vdots\\
X_p^{(t+1)} &\sim f_p(x_p | x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)}).
\end{aligned}$$

$\Rightarrow$ Zusammensetzung von p Metropolis-Hastings-Algorithmen mit Akzeptanzwahrscheinlichkeiten 1 und

$$\begin{aligned}
q_i(\tilde{\mathbf{x}} | \mathbf{x}^{(t)}) = {} & \delta_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)}\\
& \times (\tilde{x}_1, \ldots, \tilde{x}_{i-1}, \tilde{x}_{i+1}, \ldots, \tilde{x}_p) f_i(\tilde{x}_i | x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t)}, \ldots, x_p^{(t)}).
\end{aligned}$$

Satz (Konvergenz des Gibbs–Samplers)

Ist die vom Gibbs–Sampler erzeugte Markov–Kette $(\mathbf{X}^{(t)})$ ergodisch, so ist f die Dichte der stationären Verteilung und es gilt für jede Startverteilung μ:

$$\lim_{n\to\infty} \left\| \int \mathcal{K}^n(x, \cdot)\mu(dx) - f \right\|_{TV} = 0.$$

Ergodizität von $(\mathbf{X}^{(t)})$ kann charakterisiert werden.

Beispiel

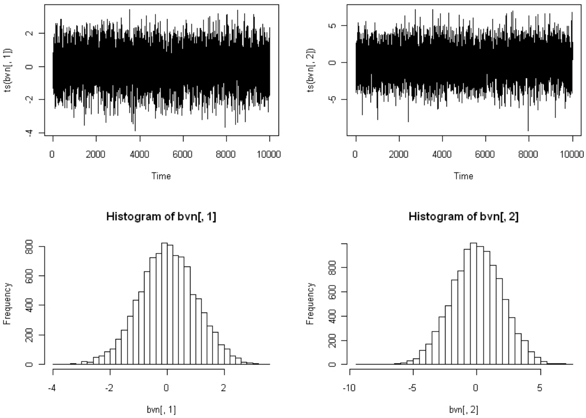
Bivariate Normalverteilung:

$$(X, Y)^T \sim \mathbf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

Die bedingten univariaten Verteilungen sind gegeben durch

$$\begin{aligned} \mathcal{L}(X|Y = y) &= \mathbf{N}\left(\rho y, (1-\rho^2)\sigma_1^2\right), \\ \mathcal{L}(Y|X = x) &= \mathbf{N}\left(\rho x, (1-\rho^2)\sigma_2^2\right). \end{aligned}$$

R–Code: Gibbs–Sampler bivariate Normalverteilung



R–Code: Gibbs–Sampler bivariate Normalverteilung