SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Gibbsovo uzorkovanje

Filip Boltužić
Voditelj: *Prof. dr. sc. Bojana Dalbelo-Bašić*

Zagreb, rujan 2013.

SADRŽAJ

1. Uvod

2. Povijest

Stuart i Donald Geman su prvi puta opisali postupak Gibbsovog uzorkovanja (Geman i Geman, 1984). Braća Geman bavili su se izradom modela za analizu slike. Gibbsovo uzorkovanje u njihovom radu bio je poseban slučaj Metropolis-Hastings algoritma (Metropolis et al., 1953). (Gelfand i Smith, 1990) su pokazali potencijalne primjene Gibbsovog uzorkovanja prilikom rješavanja velikog broja statističkih problema.

Metoda Gibbsovog uzorkovanja kasnije se koristila za uzorkovanje skupova podataka s velikim brojem varijabli. Gibbsovo uzorkovanje najčešće se koristi kada su zadovoljena dva preduvjeta. Prvi preduvjet nalaže da zajednička distribucija (engl. *joint distribution*) nije eksplicitno poznata ili je zahtjevno izravno uzorkovati iz zajedničke distribucije. Drugi preduvjet je poznata uvjetna distribucija svake varijable te mogućnost relativno jednostavnog uzorkovanja iz uvjetnih distribucija.

Gibbsov algoritam uzorkovanja za svaku varijablu generira nizove iz pripadajućih uvjetnih distribucija.

3. Metoda Gibbsovog uzorkovanja

Zajednička distribucija (engl. *joint distribution*) definirana je jednadžbom:

$$f(x, y_1, y_2, \ldots, y_p). \tag{3.1}$$

Potrebno je izračunati svojstva marginalne distribucije (engl. *marginal distribution*)

$$f(x) = \int \ldots \int f(x, y_1, y_2, \ldots, y_p) dy_1 dy_2 \ldots dy_p \tag{3.2}$$

kao što su srednja vrijednost (engl. *mean*) ili standardna devijacija (engl. *standard deviation*). Analitičkim izračunom integrala 3.2 dobije se $f(x)$, nakon čega je moguće izračunati željena svojstva. Analitički (ili numerički) izračun integrala može biti izuzetno složen. Gibbsovo uzorkovanje je alternativan način računjanja marginalne distribucije $f(x)$.

Gibbsovim uzorkovanjem generiraju se uzorci $X_1, ..., X_m$ $f(x)$ bez poznate funkcije $f(x)$. Generiranjem dovoljno velikog uzorka, moguće je izračunati svojstva, kao što su srednja vrijednost ili standardna devijacija, funkcije $f(x)$ s određenom preciznošću.

3.1. Dvodimenzionalni slučaj

Prvi primjer Gibbsovog uzorkovanja bit će objašnjen za dvodimenzionalni slučaj. Gibbsovim uzorkovanjem se za par slučajnih varijabli (X, Y) želi dobiti $f(x)$. Poznate su uvjetne distribucije $f(x|y)$ i $f(y|x)$. Generira se Gibbsova sekvenca:

$$Y_0^{'}, X_0^{'}, Y_1^{'}, X_1^{'}, \ldots, Y_k^{'}, X_k^{'}. \tag{3.3}$$

Postavlja je inicijalna vrijednost $Y_0^{'} = y_0^{'}$, dok se sve ostale vrijednosti generiraju prema

$$\begin{aligned} X_j^{'} &\sim f(x|Y_j^{'} = y_j^{'}) \\ Y_{j+1}^{'} &\sim f(y|X_j^{'} = x_j^{'}). \end{aligned} \tag{3.4}$$

Generiranje niza (3.3) prema formuli (3.4) naziva se **Gibbsovo uzorkovanje**. (Gelfand i Smith, 1990) su predložili generiranje m nezavisnih Gibbsovih sekvenci duljine k. Posljednje vrijednosti $X_k^{'}$ svake od m sekvenci se potom koriste za aproksimaciju $f(x)$. Ako je k dovoljno velik, uzorak $X^{'}$ je nezavisna i jednako distribuirana varijabla (engl. *independent and identically distributed*) kao i inicijalna nasumična varijabla X. Primjere dvodimenzionalnog slučaja Gibbsovog uzorkovanja pokazali su (Casella i George, 1992).

3.1.1. Primjer 1

Primjer zajedničke distribucije nasumičnih varijabli X i Y:

$$f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$
$$x = 0, 1, \ldots, n$$
$$0 \leq y \leq 1. \tag{3.5}$$

Potrebno je izračunati svojstva marginalne distribucije $f(x)$ slučajne varijable X. Uvjetne distribucije su poznate:

$$f(x|y) = \binom{n}{k} y^k (1-y)^{n-k} \tag{3.6a}$$

$$f(y|x) = \frac{\Gamma(\alpha+n+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \tag{3.6b}$$

Generiranjem Gibbsove sekvence formulom (3.4) pomoću uvjetnih distribucija (3.6a) i (3.6b) dobivaju se $X_1, X_2, \ldots, X_m$ iz $f(x)$. Dobiveni $f(x)$ je aproksimacija pravog $f(x)$ kojeg je moguće analitički ili numerički izračunati iz zajedničke distribucije (3.5). U ovome primjeru analitičkim izračunom dobiva se da je

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}$$
$$x = 0, 1, \ldots, n. \tag{3.7}$$

Ovdje je moguće usporediti koliko je precizno Gibbsovo uzorkovanje.

3.1.2. Primjer 2

Uvjetne distribucije slučajnih varijabli X i Y su eksponencijalne distribucije

$$f(x|y) \propto ye^{-yx}, 0 < x < B < \infty$$
$$f(y|x) \propto xe^{-xy}, 0 < y < B < \infty, \tag{3.8}$$

gdje je B poznata konstanta veća od nule. Ograničenje uvjetnih distribucija na interval $(0, B)$ je dovoljan uvjet za postojanje marginalne distribucije $f(x)$.

Prosjek konačnih vrijednosti $Y_k^{'}$ i $X_k^{'}$ Gibbsovih sekvenci može poslužiti za izračun prave marginalne distribucije. Ako se generira m sekvenci Gibbsovim uzorkovanjem onda se vrijednost f(x) može aproksimirati

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x|y_i). \tag{3.9}$$

Jednadžba (3.9) je procjena gustoće. Prilikom izračuna $f(x)$ koristi se informacija o prethodnom stanju y_1, ...,y_m iz m Gibbsovih sekvenci. Procjena sadrži više informacija od procjene s vrijednostima x_1, ... x_m. Reo-Blackwell teorem sadrži dokaz (Casella i Robert, 1996).

4. Dokaz konvergencije

Potreban je dokaz da Gibbsova sekvenca (3.3) proizvodi konvergentne nizove za nasumična varijabla distribucije $f(x)$.

X i Y su nasumične varijable, sa zajedničkom raspodjelom

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix} \tag{4.1}$$

Marginalna distribucija x je

$$f_x = \begin{bmatrix} f_x(0) & f_x(1) \end{bmatrix} = \begin{bmatrix} p_1 + p_3 & p_2 + p_4. \end{bmatrix} \tag{4.2}$$

Prema tome, uvjetne distribucije $X|Y = y$ i $Y|X = x$ iznose:

$$A_{y|x} = \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix}, A_{x|y} = \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix} \tag{4.3}$$

Dobivene matrice sliče matricama prijelaza karakterističnim za Markovljeve lance (Gilks et al., 1996). Generiranje Gibbsove sekvence (3.3) zahtjeva uvjetne distribucije, što je prikazano (4.3). U ovom slučaju Gibbsova sekvenca bit će niz nula i jedinica. Prema (3.4) potrebno je povezati uvjetne distribucije za dobivanje koraka Gibbsove sekvence, iz čega nastaje

$$P(X_1^{'} = x_1 | X_0^{'} = x_0) = \sum_y P(X_1^{'} | Y_1^{'} = y_1) \cdot P(Y_1^{'} = y_1 | X_0^{'} = x_0) \tag{4.4}$$

, u matričnom obliku

$$A_{x|x} = A_{y|x} A x|y. \tag{4.5}$$

Vrijedi:

$$f_k = f_0 A_{x|x}^k = (f_0 A_{x|x}^{k-1}) A_{x|x} = f_{k-1} A_{x|x} \tag{4.6}$$

Korak k Gibbsove sekvence se dobije kao $(A_{x|x}^k$. Ako su vrijednosti u $A_{x|x}$ pozitivne, onda (4.6) za bilokoju inicijalnu vjerojatnost f_0 i kada $k \to \infty$, f_k konvergiira

distribuciji f koja je stacionarna točna niza (4.6) i zadovoljava jednakost

$$fA_{x|x} = f. \tag{4.7}$$

Ako se generiranje Gibbsove sekvence zaustavi kod dovoljno velikog broja koraka k, pretpostavlja se kako je distribucija $X_k^{'}$ približno f_x.

Sve navedeno ne vrijedi samo u slučaju $2x2$, već i u općem slučaju slučajnih varijabli X i Y s n i m mogućih vrijednosti.

4.1. Matematika za slučaj dvije varijable

Dvije slučajne varijable X i Y. Poznate su uvjetne vjerojatnosti $f_{X|Y}(x|y)$ i $f_{Y|X}(y|x)$. Moguće je izračunati marginalnu distribuciju varijable X: $f_X(x)$, kao i zajedničku distribuciju X i Y preko:

$$f_X(x) = \int f_{XY}(x,y)dy, \tag{4.8}$$

gdje je $f_{XY}(x,y)$ zajednička distribucija.

$$f_{XY}(x,y) = f_{X|Y}(x|y)f_Y(y) \tag{4.9}$$

$$f_X(x) = \int f_{X|Y}(x|y)f_Y(y)dy$$

$$\begin{aligned} f_X(x) &= \int f_{X|Y}(x|y) \int f_{Y|X}(y|t)f_X(t)dtdy \\ &= \int \left[\int f_{X|Y}(x|y)f_{Y|X}(y|t)dy \right] f_X(t)dt \\ &= \int h(x,t)f_X(t)dt, \end{aligned}$$

gdje je

$$h(x,t) = \left[\int f_{X|Y}(x|y)f_{Y|X}(y|t)dy \right]. \tag{4.10}$$

4.2. Slučaj s više od dvije varijable

U slučaju više od dvije varijable generiranje Gibbsove sekvence radi se uzorkovanje supstitucijom (engl. *substitution sampling*).

U slučaju dvije varijable uzorkovanje supstitucijom je uvijek isto.

Za tri slučajne varijable X, Y i Z potrebno je izračunati marginalnu distribuciju $f_X(x)$. Ako se Y i Z promatraju kao jedna varijabla moguće je jednadžbom

$$f_X(x) = \int \left[\int \int f_{X|YZ}(x|y,z) f_{YZ|X}(y,z|t) dy dz \right] f_X(t) dt \tag{4.11}$$

izračunati marginalnu distribuciju. Gibbsova sekvenca bi za j-ti korak bila:

$$\begin{aligned} X_j^{'} &\sim f(x|Y_j^{'} = y_j^{'}, Z_j^{'} = z_j^{'}) \\ Y_{j+1}^{'} &\sim f(y|X_j^{'} = x_j^{'}, Z_j^{'} = z_j^{'}) \\ Z_{j+1}^{'} &\sim f(z|X_j^{'} = z_j^{'}, Y_{j+1}^{'} = y_{j+1}^{'}) \end{aligned}$$

5. Zaključak

Zaključak.

6. Literatura

George Casella i Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

George Casella i Christian P Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

Alan E Gelfand i Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

Stuart Geman i Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Walter R Gilks, Sylvia Richardson, i David J Spiegelhalter. *Markov chain Monte Carlo in practice*, svezak 2. CRC press, 1996.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, i Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.