نمونه گیری گیبس

۱.مقدمه

تکامل روزافزون رايانه و دسترسي آسان به رايانه هاي سريع باعث شده است که روشهاي الگوريتمي همچون نمونه گيري گيبس ابزارهايي مفيد و ضروري براي حل مسائل آماري باشند که حل آنها به صورت تحليلي ميسر نيست. نمونه گيري گيبس یک روش الگوريتمي (مثل الگوريتم EM) است که ايده اوليه آن توسط متروپليس و تلر در سال ۱۹۵۳ ارائه و سپس با مقاله اي که گيمان و گيمان {۴} در مورد مدلهاي پردازش تصوير ارائه دادند وارد مرحله جديدي شد.

نمونه گيري گيبس اساسا مبتني بر خاصيت ماركوفي فرايندهاي تصادفي و روشهاي مونت كارلوست به طوري که گاهي نمونه هاي توليد شده با استفاده از اين روش را نمونه مونت كارلو نيز مي نامند.

۲.نمونه گیری گیبس چیست؟

نمونه گيري گيبس را مي توان یک ابزار شبيه سازي براي به دست آوردن نمونه هاي تصادفي از تابع توزيع توام نرمال نشده دانست. اين نمونه ها نمونه هايي را نيز از توزيعهاي حاشيه اي مربوط به توزيع توام توليد مي كنند.

در بسياري از مسائل آماري با مجموعه اي از متغيرهاي تصادفي سروكار داريم که علاقه مند به توزيعهاي حاشيه اي آنها يا اطلاعاتي در مورد آنها هستيم. در چنين مواقعي اگر توزيع توام نرمال شده اين متغيرها موجود باشد مي توانيم چگاليهاي حاشيه اي را به طور تحليلي و دقيق به دست آوريم. غالبا در بسياري از اين مسائل توزيع توام متغيرها در عمل نامعلوم است اما مجموعه اي از توزيعهاي شرطي وجود دارد. اگر مجموعه مناسبي از چگاليهاي شرطي موجود باشد مي توان با توسل به روشهاي مونت كارلو و به كارگيري تكنيكهاي نمونه گيري مجدد مانند نمونه گيري جانشين ساز(جايگشتي) و يا نمونه گيري گيبس یک نمونه موثر از چگالي توام و احتمالا حاشيه اي مورد نظر را توليد كرده و سپس به وسيله نمونه توليد شده تابع چگالي و يا خصيصه هاي مربوط به آن را با درجه اطمينان دلخواهي محاسبه كرد.

نمونه گيري گيبس در واقع راهي براي توليد نمونه تصادفي از یک تابع چگالي حاشيه اي است بدون آن که نيازي به محاسبه خود تابع چگالي باشد. بيشترين كاربرد نمونه گيري گيبس استفاده از آن براي به دست آوردن توزيعهاي پسين در مدلهاي بيزي است. علاوه بر اين در روشهاي كلاسيک نيز(براي محاسبه توابع درستنمايي و تعيين برآوردهاي MLE) استفاده از اين روش بسيار سودمند است.

فرض كنيد توزيع توام $f(x, y_1, \ldots, y_p)$ معلوم باشد و بخواهيم برخي ازمشخصات توزيع حاشيه اي را مثل ميانگين و واريانس به دست آوريم. احتمالا به طور معمول بايد $f(x)$ را محاسبه كرده و با استفاده از آن مشخصات مورد نظر را محاسبه كنيم.

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_p) dy_1 \ldots dy_p \qquad (۱)$$

در بسياري از حالتها محاسبه انتگرال (۱) مستقيما يا با کمک آناليز عددي بسيار مشكل است. در چنين حالتهايي نمونه گيري گيبس راه حل ديگري براي به دست آوردن $f(x)$ ارائه مي كند.

به جاي محاسبه يا تقريب $f(x)$ به طور مستقيم با استفاده از نمونه گيري گيبس مي توان یک نمونه موثر $x_1, \ldots, x_m$ از توزيع $f(x)$ را بدون نياز به $f(x)$ توليد كرد. با شبيه سازي یک نمونه به اندازه كافي بزرگ ميانگين و واريانس يا هر مشخصه ديگر $f(x)$ را مي توان با درجه اطمينان دلخواهي به دست آورد. واقعيت مهم اين است که نتايج نهايي محاسبات هرچند که بر اساس شبيه سازي است ولي برآوردهايي براي كميتهاي جامعه هستند. به عنوان مثال براي محاسبه ميانگين $f(x)$ از

$$\frac{1}{m}\sum_{i=1}^{m} x_i$$

و اين حقيقت که با احتمال یک داريم:

$$\frac{1}{m}\sum_{i=1}^{m} x_i \to E(X) \qquad (۲)$$

استفاده مي كنيم. بنابراين وقتي كه m به اندازه كافي بزرگ اختيار شود هر مشخصه جامعه حتي خود چگالي را مي توان با درجهاطمينان دلخواه به دست آورد. براي درک نمونه گيري گيبس ابتدا حالت دو متغيره را بررسي مي كنيم.

۲-۱ حالت دو متغیره

دو متغير تصادفي XوY را در نظر بگيريد. در نمونه گيري گيبس براي توليد يک نمونه از چگالي $f(x)$ از چگاليهاي شرطي $f(x|y)$ و$f(y|x)$ (كه اغلب در مدلهاي آماري معلوم اند) به جاي خود $f(x)$ استفاده مي شود. اين كار با توليد يک دنباله از متغيرهاي تصادفي

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \dots, Y_K', X_K' \qquad (۳)$$

كه آن را دنباله گيبس مي نامند انجام مي شود. براي توليد دنباله گيبس (۳) ابتدا مقدار اوليه $Y_0' = y_0'$ را تعيين كرده و سپس (۳) را به طور مكرر و متناوب با استفاده از

$$X_i' \sim f(x \mid Y_i' = y_i')$$

(۴)

$$Y_{i+1}' \sim f(y \mid X_i' = x_i')$$

توليد مي كنيم. تحت شرايط كلي به طور منطقي ثابت مي شود كه وقتي $k \to \infty$ توزيع X_k' به $f(x)$ توزيع حاشيه اي X همگرا مي شود. بنابراين براي K به اندازه كافي بزرگ مشاهدات نهاي $X_0' = x_0'$ در (۳) واقعا يک نمونه موثر از نقاط $f(x)$ است.براي مثال اسميت{۲}m دنباله گيبس مستقل را با طول K توليد كرد و سپس مقادير نهايي X_k' را از هريک از دنباله ها به كاربرد. در اين روش اگر K به اندازه كافي بزرگ انتخاب شود يک نمونه i.i.d از $f(x)$ به دست مي آيد. براي روشن شدن مطلب روش بالا را در مثالهاي زير توضيح مي دهيم.

مثال ۱ – فرض كنيدX وY دومتغير تصادفي با تابع چگالي توام زير باشند:

$$f(x,y) \propto \binom{n}{x} y^{x+\alpha+1}(1-y)^{n-x+\beta-1} \qquad (۵)$$

$$0 \le y \le 1 \,;\, x = 0,1,\dots,n$$

همچنين فرض كنيد علاقه مند به محاسبه برخي از مشخصات $f(x)$ (توزيع حاشيه اي X) باشيم. با استفاده از نمونه گيري گيبس يک نمونه از توزيع حاشيه ايX را توليد مي كنيم. از(۵) (بدون اين كه به n, β, α بستگي داشته باشد) نتيجه مي شود كه

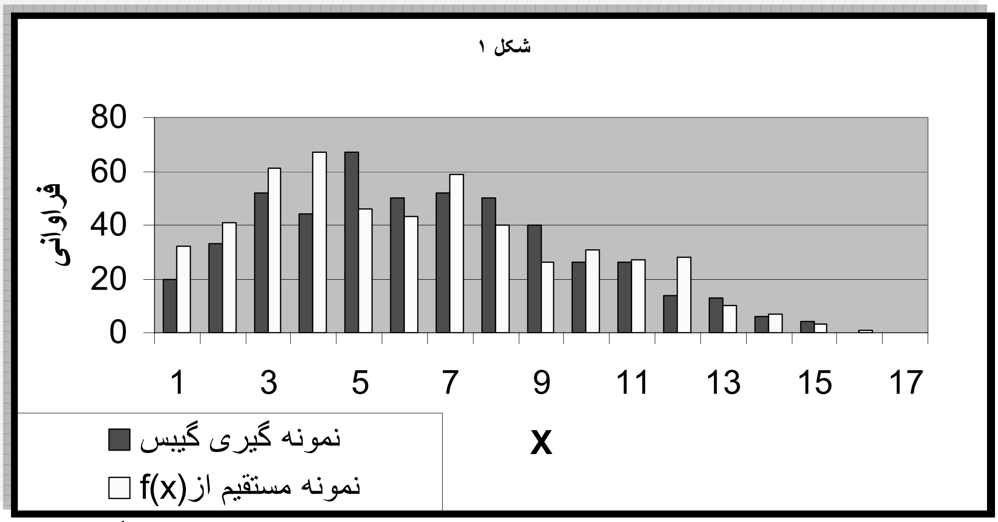
$$f(x|y) \sim Binomial(n, y)$$

$$f(y|x) \sim Beta(x+\alpha, n-x+\beta) \qquad (۶)$$

با به كارگيري طرح مكرر(۴) براي توزيعهاي (۶)مي توانيم يک نمونه $x_1, \dots, x_m$ را از $f(x)$ توليد كرده و سپس با استفاده از آن مشخصات مورد نظر را برآورد كنيم. ممكن است خواننده عنوان كند كه در اين مثال نيازي به نمونه گيري گيبس نيست زيرا مي توان $f(x)$ را از رابطه (۵) به صورت توزيع بتا-دوجمله اي به دست آورد و سپس مشخصات مورد نياز $f(x)$ را مستقيما با استفاده از آن و بدون آن كه خود را باتوزيعهاي شرطي درگير كنيم محاسبه كرد. به هر حال اين مثال مثالي ساده براي نشان دادن اين موضوع است كه:

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)} \qquad (۷)$$

$$; x = 0,1,\dots,n$$



شکل (۱) بافتنگار دو نمونه $x_1, x_2, ..., x_m$ را با $m = 500$ و $n = 16$ و $\beta = 4$ و $\alpha = 2$ که به دو روش از چگالي $f(x)$ توليد شده اند نشان مي دهد. در یک روش این نمونه مستقيما از چگالي(۷) توليد شده و در روش ديگر اين نمونه با استفاده از چگاليهاي شرطي (۶) و طرح نمونه گيري گيبس با $k = 15$ توليد شده است. تشابه دو بافتنگار در شکل(۱) تاييدي بر اين ادعاست که طرح نمونه گيري گيبس واقعا متغيرهايي از توزيع حاشيه اي را توليد مي کند. اين دو نمونه به وسيله برنامه اي که با نرم افزار s-plus نوشته شده است به دست آمده اند.
نکته اي که بايد به آن توجه داشته باشيم اين است که نمونه گيري گيبس نيازي به توزيع توام که با استفاده از رابطه

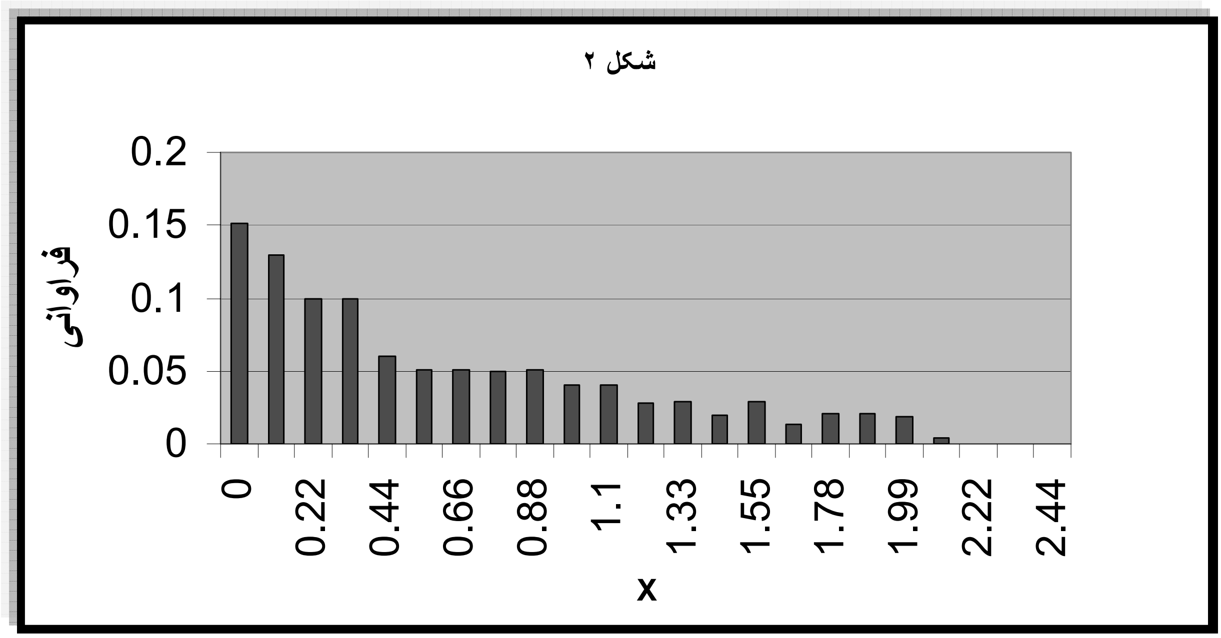
$$f(x) = \frac{f(x, y)}{f(y \mid x)}$$

$f(x)$ را محاسبه کند ندارد. مثال بعد نشان مي دهد که نمونه گيري گيبس براي وقتي که $f(x)$ و $f(y)$ يا $f(x, y)$ قابل محاسبه نباشند ضروري است.

مثال۲– فرض کنیدXوY داراي توزيعهاي شرطي نمايي محدود شده به بازه (0,B) باشند يعني

$$f(x \mid y) \propto y e^{-yx}; 0 < x < B < \infty \qquad (۸)$$
$$f(y \mid x) \propto x e^{-xy}; 0 < y < B < \infty$$

که در آن B يک ثابت معلوم است. محدوديت به بازه(0,B) ما را مطمئن مي کند که توزيع حاشيه اي $f(x)$ وجود دارد. اگر B متناهي نباشد آنگاه چگاليهاي(۸) يک زوج معتبر از چگاليهاي شرطي نيست. به اين مفهوم که چگالي توام $f(x, y)$ وجود ندارد و در اين صورت دنباله گيبس همگرا نخواهد شد. شرايط تعيين چگاليهاي حاشيه اي و همچنين جزئيات مربوط به شرايط همگرايي توسط کازلين و شرويش در سال ۱۹۹۰ مورد بررسي قرار گرفته است.
در اين مثال توزيع حاشيه اي به سادگي قابل محاسبه نيست ولي با استفاده از نمونه گيري گيبس براي شرطهاي (۸) هر مشخصه $f(x)$ را مي توان به دست آورد. شکل (۲) بافتنگار يک نمونه $m = 500$ تايي از $f(x)$ را که با استفاده از مقادير نهايي دنباله گيبس به طول $k = 15$ براي $B = 5$ به دست آمده نشان مي دهد.



۲–۲بر آورد چگالیهای حاشیه ای

نمونه گيري گيبس را مي توان براي برآورد خود چگالي حاشيه اي نيز به كار برد. در دنباله گيبس (۳) $X'_k = x'_k$ یک تحقق از $f(x)$ يعني $x_1, x_2, ..., x_m$ از توزیع $f(x)$ و $Y'_k = y'_k$ یک تحقق از $f(y)$ يعني $y_1, y_2, ..., y_m$ از توزیع $f(y)$ هستند. علاوه بر اين ميانگين چگاليهاي شرطي $f(x \mid Y_k = y_k)$ به $f(x)$ ميل مي كند. بنابراين براي محاسبه $f(x)$ كافي است ميانگين مقادير نهايي چگاليهاي شرطي دنباله گيبس را محاسبه كنيم يعني $f(x)$ را با استفاده از

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f(x \mid y_i) \qquad (۹)$$

برآورد نماييم. در رابطه(۹) $y_1, y_2, ..., y_m$ مقادير نهايي به دست آمده دنباله گيبس هستند. پشتوانه فرمول(۹) اين است كه

$$E[f(x \mid y)] = \int f(x \mid y) f(y) dy = f(x) \quad (۱۰)$$

روش برآورد چگالي (۹) را مي توان براي توزيعهاي گسسته نيز به كار برد. مثلا اگر مقادير $y_1, y_2, ..., y_m$ نهاييy درm نمونه گيبس براي چگاليهاي شرطي مثال(۱) باشند مي توان متناظر با (۹) احتمالهاي حاشيه ايX را با استفاده از

$$\hat{P}(X = x) = \frac{1}{m}\sum_{i=1}^{m} P(X = x \mid Y_i = y_i) \qquad (۱۱)$$

برآورد كرد. اين روش را براي برآورد چگالي $f(x)$ در حالت سه متغيره در مثال ۳ به كار مي بريم.

۳–۲ حالت سه متغیره

واقعيتي كه نمونه گيري گيبس از آن استفاده مي كند اين است كه اطلاعات نهفته در توزيعهاي شرطي براي تعيين توزيع توام (درصورت وجود) كافي است. در حالت دو متغيره نمونه گيري گيبس براي به دست آوردن چگاليهاي حاشيه اي از چگاليهاي شرطي به سرعت پيش مي رود. اما در حالت چند متغيره روابط ميان چگاليهاي شرطي و حاشيه اي و توام بسيار پيچيده مي شوند. به عنوان مثال رابطه

حاشيه اي× شرطي = توام

براي تمام چگاليهاي شرطي و چگاليهاي حاشيه اي به دست نمي آيد. بدين معني كه ممكن است راههاي متعددي براي ايجاد يک چگالي حاشيه اي از چگاليهاي شرطي وجود داشته باشد و بتوان مجموعه هاي مختلفي از توزيعهاي شرطي را براي محاسبه توزيع حاشيه اي مورد نظر به كار برد.

تعريف- فرض كنيد $U_1, U_2, ..., U_k$ k متغير تصادفي باشند. چگاليهاي شرطي هر متغير به شرط(k-1) متغير ديگر را يک چگالي شرطي كامل يا شرطي يک متغيره مي ناميم. بنابراين براي اين k متغيره k چگالي شرطي كامل

$$f_{U_i|U_{i1},...U_{ik-1}}(u \mid u_1, ..., u_{k-1}); i = 1,2,...k$$

$$i \neq j_t, t \in \{1, ..., k-1\}$$

وجود دارند.

در نمونه گيري گيبس هميشه مجموعه تمام چگاليهاي شرطي كامل براي تعريف تكرارها به كار برده مي شوند. بساگ در سال ۱۹۷۴ نشان داد كه اين مجموعه از چگاليهاي شرطي براي تعيين توزيع توام(و در نتيجه براي تعيين هر توزيع حاشيه اي) به طور يكتا كافي است و اين كوچكترين مجموعه لازم از چگاليهاي شرطي براي تعريف تكرارهاست.

فرض كنيد براي سه متغير تصادفي X وY وZ چگاليهاي شرطي كامل يعني $f_{X|YZ}$ و $f_{Y|XZ}$ و $f_{Z|XY}$ وجود داشته باشند. نمونه گيري گيبس در هر تكرار از اين چگاليهاي شرطي نمونه گيري مي كند يعني i-امين تكرار در نمونه گيري گيبس به صورت زير است:

$$X'_j \sim f(x \mid Y'_j = y'_j, Z'_j = z'_j)$$
$$Y'_{j+1} \sim f(y \mid X'_j = x'_j, Z'_j = z'_j) \qquad (۱۲)$$
$$Z'_{j+1} \sim f(z \mid X'_j = x'_j, Y'_{j+1} = y'_{j+1})$$

روش مكرر(۱۲) يک نمونه

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, Y'_2, Z'_2, X'_2 \qquad (۱۳)$$

را با خاصيت كه براي مقادير بزرگ k, $X'_k = x'_k$ يک نمونه موثر از $f(x)$ است توليد مي كند.

مثال۳– در توزيع مثال (۱), N را يک متغير تصادفي پواسن با پارامتر λ در نظر مي گيريم و لذا توزيع توام به صورت

$$f(x, y, n) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} e^{-\lambda} \frac{\lambda^n}{n!} \quad (۱۴)$$

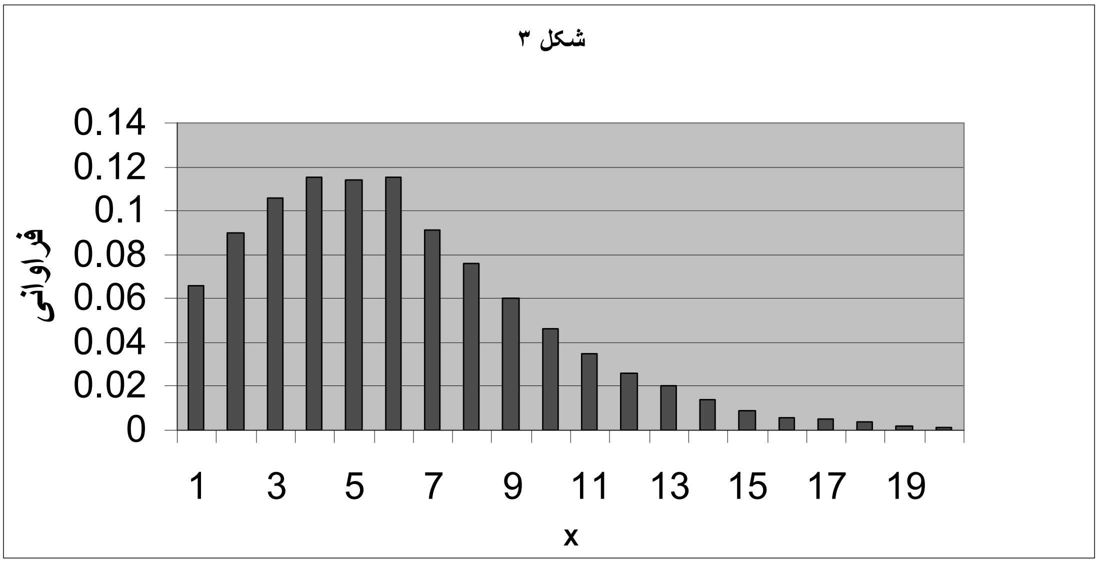
نوشته مي شود. فرض كنيد علاقه مند به توزيع حاشيه اي X باشيم. برخلاف مثال(۱) در اين جا نمي توان توزيع حاشيه اي X را به سرعت محاسبه كرد. با توجه به (۱۲) سه چگالي شرطي كامل را نياز داريم. اين چگاليها به صورت زير به دست مي آيند:

$$f(x \mid y, n) \sim Binomial(n, y)$$
$$f(y \mid x, n) \sim Beta(x+\alpha, n-x+\beta) \qquad (۱۵)$$
$$f(n \mid x, y) \sim e^{-(1-y)} \frac{[(1-y)\lambda]^{n-x}}{(n-x)!} \lambda$$
$$n = x, x+1, ...$$

اگر m نمونه گيبس(۱۳) را توليد كنيم مي توانيم با استفاده از رابطه اي شبيه به رابطه(۱۱) چگالي X را به صورت زير برآورد كنيم:

$$\hat{P}(X = x) =$$
$$\frac{1}{m^2} \sum_{i=1}^{n} \sum_{j=1}^{m} P(X = x \mid Y_i = y_i, N = n_j) \qquad (۱۶)$$

در رابطه(۱۶) $x_1, ..., x_m$ و $y_1, ..., y_m$ مقادير به دست آمده از m نمونه گيبس اند. شكل(۳) برآورد احتمالهاي توزيع حاشيه اي X را كه با استفاده از (۱۶) بر اساس يک نمونه $m = 500$ تايي با دنباله هايي به طول $k = 10$ براي سه چگالي(۱۵) با $n = 16$ و $\alpha = 2$ و $\beta = 4$ به دست آمده است نشان مي دهد.



۲–۴ حالت چند متغیره

براي سادگي كار توزيعها را به طور كلي با براكت([]) نشان مي دهيم. مثلا توزيعهاي توام و شرطي و حاشيه اي را به صورت [X,Y] و[X|Y] و [X] نشان مي دهيم. همچنين علامت* را براي حاصلضرب چگاليها به كار مي بريم. حال با استفاده از اين نماد گذاري روند نمونه گيري گيبس را مي توان براي k متغير به صورت زير بيان كرد.

فرض كنيد, $U_1,U_2,...,U_k$ k متغير تصادفي باشند كه چگاليهاي شرطي كامل آنها وجود دارند. همچنين فرض كنيد $U^{(0)}{}_1,...,U^{(0)}{}_k$ يك مجموعه آغازي دلخواه باشد. در اين صورت $U^{(1)}{}_1,...,U^{(1)}{}_k$را به ترتيب از

$U_1^{(1)} \sim [U_1 \,|\, U_2{}^{(0)}, U_3^{(0)},...,U_k{}^{(0)}]$

$U_2^{(1)} \sim [U_2 \,|\, U_1{}^{(1)}, U_3^{(0)},...,U_k{}^{(0)}]$

$U_3^{(1)} \sim [U_3 \,|\, U_1{}^{(1)}, U_3^{(1)}, U_4^{(0)},...,U_k{}^{(0)}]$

$\vdots$

$U_k^{(1)} \sim [U_k \,|\, U_1{}^{(1)},...,U_{k-1}{}^{(1)}]$

توليد مي كنيم. با اين روند هريک از متغيرها به نوبت در يک چرخه ظاهر شده و هرچرخه شامل k متغير تصادفي توليد شده است. پس ازi تكرار از اين چرخه ها $(U_1{}^{(i)},...,U_k{}^{(i)})$ توليد مي شود. گيمان و گيمان در سال ۱۹۸۴ ثابت كردند كه وقتي $i \to \infty$ داريم:

$(U_1{}^{(i)},...,U_k{}^{(i)}) \xrightarrow{d} [U_1,...,U_k]$

$U_s{}^{(i)} \xrightarrow{d} U_s \sim [U_s]$ و همچنين

الگوریتم نمونه گیری گیبس

الگوريتم نمونه گيري گيبس براي متغير تصادفي $(X_1,..., X_p)$ با چگالي $f(x) = cg(x)$ به صورت زير است. ابتدا بررسي كنيد كه g مثل چگالي كه بتوان آن را راحت شبيه سازي كرد هست يا خير. سپس مراحل زير را طي كنيد:

۱. يک مقدار شروع مانند $X_2^{(0)},..., X_p^{(0)}$ براي $X_2,..., X_p$ انتخاب كنيد و قرار دهيد $i = 0$.

۲. مقدار جديد $x_1^{(i+1)}$ از توزيع شرطي X_1 به شرط; $X_p = x_p^{(i)},..., X_2 = x_2^{(i)}, X_2 = x_2^{(i)}$ را شبيه سازي كنيد.

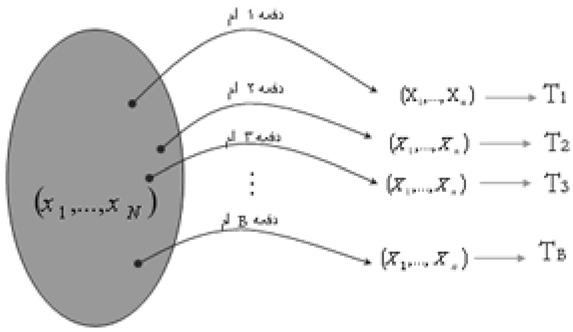
۳. مقدار جديد $x_2^{(i+1)}$ را از توزيع شرطي X_2 به شرط; $X_p = x_p^{(i)},..., X_3 = x_3^{(i)}, X_1 = x_1^{(i+1)}$ شبيه سازي كنيد.

۴. $p+1$ مقدار جديد $x_p^{(i+1)}$ را از توزيع شرطي X_p به شرط $X_{p-1} = x_{p-1}^{(i+1)},..., X_1 = x_1^{(i+1)}$ شبيه سازي كنيد.

۵. $p+2$.$i+1$ را جايگزين i كنيد و به مرحله۲ بازگرديد.

دنباله p تايي هاي $(X_1,...,X_p)$ يک زنجير ماركوف تشكيل مي دهند. توزيع ماناي اين زنجير ماركوف داراي چگالي f و توزيع يک تكرار پس از مراحل بسيار تقريبا همان توزيع مانا را خواهد داشت.

محاسبه اریبی و خطای معیار



فرض كنيد X1...Xn متغيرهاي تصادفي مستقل و همتوزيع (i.i.d) با تابع توزيع تجمعي معلوم F وT آماره (برآوردگر) مورد نظر باشد. با استفاده از شبيه سازي مونت كارلو مي توان توزيع، اريبي، خطاي معيار و ... T را بصورت مراحل زير تقريب كرد .

مرحله ۱- نمونه تصادفي مستقل و همتوزيع از توزيع F بدست مي آيد.

مرحله ۲ - آماره T بر روي نمونه مرحله ۱ بصورت T1...Tn) T) مي گردد.

مرحله ۳ - مراحل ۱ و ۲ را B بار تكرار مي شود .

براي B هاي بزرگ تابع توزيع، اريبي و خطاي معيار T به صورت زير است :

$$F_T(t) \cong \frac{1}{B}\sum_{i=1}^{B} I(T_i \le t)$$

$$bias(T) \cong \frac{1}{B}\sum_{i=1}^{B} T_i - \theta$$

$$SE(t) \cong \left[\frac{1}{B}\sum_{i=1}^{B}\left(T_i - \frac{1}{B}\sum_{j=1}^{B} T_j\right)^2\right]^{\frac{1}{2}}$$

مثال کاربردی

به منظور درك بهتر شبيه سازي مونت كارلو مثال ارائه مي گردد.

مثال- در يک فرايند تحقيقاتي به منظور بدست آوردن وزن تقريبي محصول هندوانه يک مزرعه با جامعه اي نامحدود روبرو هستيم ولي اين نكته را مي دانيم كه وزن هندوانه ها از توزيع نرمال با متوسط وزن ۶ كيلو گرم و انحراف معيار۳ كيلوگرم پيروي مي كند .

اكنون سه مرحله الگوريتم مونت كارلو را در اينجا اجرا مي شود :

مرحله۱- نمونه اي تصادفي به حجم ۲۵ از توزیع نرمال با میانگین ۶ و واریانس ۹ با استفاده از نرم افزار "اس پلاس" تولید مي شود.به عنوان مثال ۶.۸۸ و ۹.۰۸ و۹.۵۹ و۵.۸۲ و ۲.۳۰ و ۳.۷۵ و ۶.۷۹ و ۱۴.۸۷ و ۱.۷۲و ۵.۳۳ و۴.۷۳ و۱۴.۶۳ ۵.۲۹ و۴.۰۹ و۵.۸۵ و۸.۱۶ و۵.۷۷ و۶.۵۲ و۲.۶۷ و۲.۹۲ و۰.۲۴- و ۳.۴۲و ۱.۰۷و ۴.۴۹ و۵.۷۵)

مرحله۲ - آماره T بر روي نمونه مرحله ۱ بصورت T (T1...Tn) مي گردد. که در اینجا به عنوان مثال میانگین را برآورد کرده :

$$\bar{X} = 0.65$$

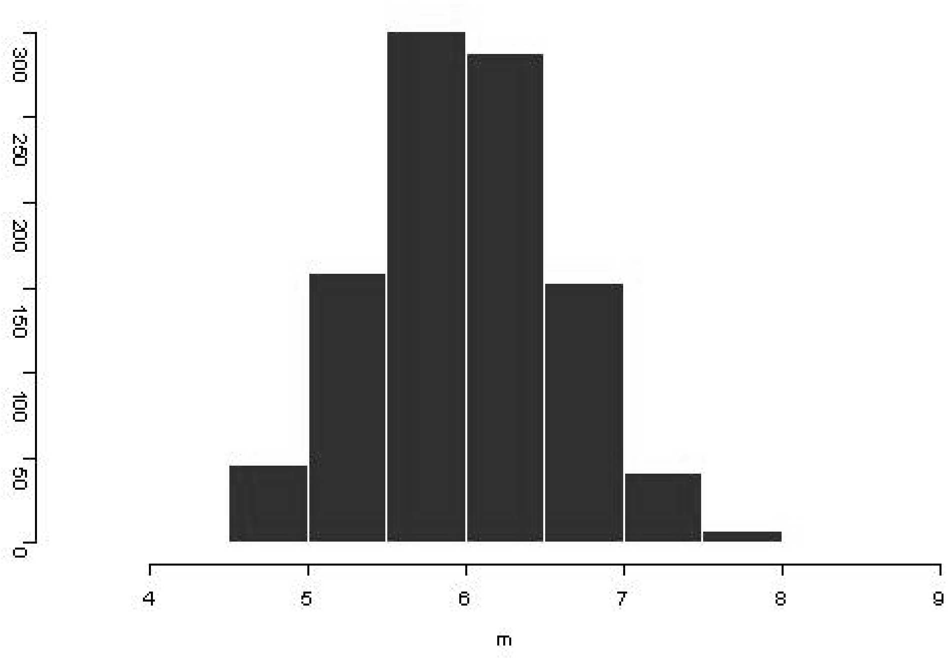
مرحله ۳- مراحل ۱ و ۲ را 10000بار تکرار مي شود و

$$\bar{X}_1, ..., \bar{X}_{10000}$$

بدست مي آید. که اریبي، خطاي معیار و توزیع برآوردگر بالا به صورت زیربرآورد مي شود :

$$bias(\bar{X}) = 0.000$$
$$SE(\bar{X}) = 0.611$$

بافتنگار زیر توزیع X را بر اساس موارد بالا نشان مي دهد :



پیاده سازی مثالی از نمونه گیری گیبس در نرم افزار R و همچنین C

Implementation in R

A function for the Gibbs sampler .

```
gibbs<-function (n, rho)
{
        mat <- matrix(ncol = 2, nrow = n)
```

```
        x <- 0
        y <- 0
        mat[1, ] <- c(x, y)
        for (i in 2:n) {
                x <- rnorm(1, rho * y, sqrt(1 - rho^2))
                y <- rnorm(1, rho * x, sqrt(1 - rho^2))
                mat[i, ] <- c(x, y)
        }
        mat
}
```

A matrix for the results is created, then the chain is initialised at (0,0). The main loop then successively samples from the full conditionals, storing the results in the matrix. We can test this as follows:

```
bvn<-gibbs(10000,0.98)
par(mfrow=c(3,2))
plot(bvn,col=1:10000)
plot(bvn,type="l")
plot(ts(bvn[,1]))
plot(ts(bvn[,2]))
hist(bvn[,1],40)
hist(bvn[,2],40)
par(mfrow=c(1,1))
```

With a bit of luck, this will give results which look very similar to those obtained earlier, apart from the time series plots of the marginals, which show distinct autocorrelation between successive values.

Implementation in C

Of course, Gibbs samplers are Markov chains, which cannot be neatly vectorised in languages like R. Consequently, the main loop of a Gibbs sampler is best re-coded in a compiled language such as C. Here is a main loop for a Gibbs sampler for this problem.

```
int main(int argc, char *argv[])
{
  long n,i;
  double x,y,rho,sd;
  gsl_rng *r = gsl_rng_alloc(gsl_rng_mt19937);
  n=(long) atoi(argv[1]);
  rho=(double) atof(argv[2]);
  sd=sqrt(1-rho*rho);
  x=0;y=0;
  printf(" %3.3f %3.3f \n",x,y);
  for (i=1;i<n;i++)
    {
      x=rho*y+gsl_ran_gaussian(r,sd);
      y=rho*x+gsl_ran_gaussian(r,sd);
      printf(" %3.3f %3.3f \n",x,y);
    }
  return(0);
}
```

. The C Code can be tested by re-directing output to a file (`bvn.dat`), then reading and analysing from R:

```
bvn_matrix(scan("bvn.dat"),ncol=2,byrow=T)
par(mfrow=c(3,2))
plot(bvn,col=1:10000)
plot(bvn,type="l")
plot(ts(bvn[,1]))
plot(ts(bvn[,2]))
hist(bvn[,1],40)
hist(bvn[,2],40)
par(mfrow=c(1,1))
```

همگرایی احتمالهای انتقال در زنجیر مارکوف برای اثبات موجه بودن نمونه گیری گیبس

زنجير ماركوفي با سه حالت و ماتريس انتقال زير را در نظر بگيريد:

$$P=\begin{pmatrix}0.2 & 0.7 & 0.1\\ 0.8 & 0.1 & 0.1\\ 0.3 & 0.3 & 0.4\end{pmatrix}$$

ديديم كه ماتريس احتمال انتقال m مرحله اي براي چنين زنجير ماركوفي به صورت P^m است. اين ماتريس احتمالها ي انتقال بين دو حالت در m مرحله از زنجير ماركوف را فراهم مي كند. مثلا با $m=20$ و ماتريس P مذكور در بالا داريم:

$$P^{20}=\begin{pmatrix}0.4553 & 0.4018 & 0.1429\\ 0.4553 & 0.4018 & 0.1429\\ 0.4553 & 0.4018 & 0.1429\end{pmatrix}$$

توجه مي كنيم كه سطرهاي ماتريس P^{20} چقدر بهم نزديك اند و حال آنكه در ماتريس P سطرها بسيار متفاوت بودند. اگر احتمال آغازين بردار w هرچه باشد wP^{20} برابر بردار $w^*=(0.4553,0.4018,0.1429)$ خواهد بود. يعني توزيع حالت آغازين هرچه كه باشد توزيع مرحله ۲۰ ام زنجير w^* است. اين بردار w^* ويژگي خاصي دارد يعني در برابري $w^*P=w^*$ صدق مي كند.
آنچه كه در مثال بالا رخ داد در تمام زنجيرهاي ماركوف با ماتريس P كه داراي درايه هاي مثبت است رخ مي دهد(در واقع حتي كلي تر از اين) آنچه كه در حالت كلي رخ مي دهد آن است كه P^m وقتي كه $m\to\infty$ به ماتريسي كه سطرهاي آن با هم برابراند ميل مي كند. اين بردار سطري مشترك را با w^* نشان مي دهد و w^* و P در برابري $w^*P=w^*$ صدق مي كنند. در مثال ما P^{20} با ماتريس حدي تا چهار رقم غير قابل تميز است. بردار w^* را توزيع مانا مي نامند زيرا اين بردار داراي اين ويژگي است كه اگر w^* توزيع حالت i باشد آنگاه توزيع تمام حالتها از آنجا به بعد نيز w^* خواهد بود. حتي اگر توزيع حالت در زمان i نباشد به علت همگرايي توزيع حالت $i+m$ به w^* براي m هاي بزرگ بسيار نزديك خواهد بود. اين واقعيت را وقتي كه براي زنجير ماركوف در حالت كلي بكار مي بريم دليلي بر موجه بودن الگوريتم نمونه گيري گيبس خواهد بود. صرفنظر از توزيعي كه زنجير ماركوف با آن شروع كرده است پس از طي تعداد زيادي از مراحل توزيع حالتها تقريبا همان توزيع مانا خواهد بود. اما به راحتي مي توان بررسي كرد كه توزيعي كه درصدد شبيه سازي آن هستيم توزيع مانا است. مثلا اگر $(X_1^{(i)},X_2^{(i)})$ داراي چگالي توام f باشند آنگاه $(X_1^{(i+1)},X_2^{(i)})$ نيز داراي چگالي f و همچنين $(X_1^{(i+1)},X_2^{(i+1)})$ نيز همين چگالي f را خواهد داشت.به علت اين شباهت با زنجير ماركوف نمونه گيري گيبس(و روشهاي متعدد وابسته به آن) را غالبا زنجير ماركوف مونت كارلو مي نامند.
توجه: براي همگرايي زنجيرهاي ماركوف شرايطي لازم است. چنين نيست كه تمام زنجيرهاي ماركوف داراي توزيع مانا باشند و چنين نيست كه تمام زنجيرهاي ماركوف با توزيع مانا همگرا باشند.

مثال– نمونه ای از توزیع نرمال

فرض کنید نمونه اي نرمال با میانگین مجهول μ و دقت مجهول τ مشاهده کرده باشیم. فرض کنید از پیشین مزدوج طبیعي استفاده کرده باشیم. حاصلضرب پیشین و درستنمایي را براي سهولت ذیلا مي آوریم:

$$\varepsilon(\mu,\tau\,|\,x) \propto \tau^{\alpha_1+\frac{1}{2}-1}\exp(-\tau[\frac{1}{2}\lambda_1(\mu-\mu_1)^2+\beta_1])$$

که α_1 و β_1 و μ_1 و λ_1 به محض معلوم شدن داده ها مقادیر معلومي خواهند بود. اگر این تابع را براي مقدار ثابت τ به عنوان تابعي از μ در نظر بگیرید شبیه چگالي توزیع نرمال با میانگین μ_1 و واریانس $\frac{1}{\tau\lambda_1}$ خواهد بود. اگر آن را به عنوان تابعي از τ براي μ ثابت درنظر بگیرید شبیه چگالي توزیع گاما با پارامترهاي $\alpha_1+\frac{1}{2}$ و $\frac{\lambda_1(\mu-\mu_1)^2}{2}+\beta_1$ است. این دو توزیع را به آساني مي توانیم شبیه سازي کنیم.

وقتي که تابع $g(x_1,x_2)$ را به صورت تابعي از x_1 براي مقدار ثابت x_2 درنظر مي گیریم به چگالي شرطي X_1 به شرط $X_2=x_2$ بجز یک عامل ضربي که به x_1 بستگي ندارد مي رسیم. به طریق مشابه وقتي که $g(x_1,x_2)$ را به صورت تابعي از x_2 براي مقدار ثابت x_1 توجه مي کنیم به چگالي شرطي X_2 به شرط $X_1=x_1$ مي رسیم.

با تعیین تابع $g(x_1,x_2)$ به صورتي که موردنظر ماست الگوریتم زیر را طي مي کنیم:

۱. یک مقدار شروع براي x_2 مانند $x_2^{(0)}$ انتخاب مي کنیم و قرار مي دهیم $i=0$.

۲. مقدار جدید $x_1^{(i+1)}$ را از توزیع شرطي X_1 به شرط $X_2=x_2^{(i)}$ شبیه سازي مي کنیم.

۳. مقدار جدید $x_2^{(i+1)}$ را از توزیع شرطي X_2 به شرط $X_1=x_1^{(i+1)}$ شبیه سازي مي کنیم.

۴. i را به $i+1$ تبدیل کنید و به مرحله ۲ بازگردید.

این الگوریتم نوعا وقتي که i به مقدار به اندازه کافي بزرگ برسد پایان مي پذیرد.

پایان

منابع:

نمونه گیري گیبس/سعید زال زاده/اندیشه آماري/سال هفتم/شماره اول

احتمال و آمار/جلد اول/دگرت موریس-اسکرویش مارک/پاشا عین اله

Probability and statistics

d.j.wilkinson@ncl.ac.uk

http://www.staff.ncl.ac.uk/d.j.wilkinson