روش‌های مونت‌کارلو در استنباط بیزی

زهرا خادم بشیری
گروه آمار، دانشگاه پیام نور تهران شرق

چکیده

از آن‌جایی که آماردانان در استنباط بیزی معمولا علاقه‌مند به انجام استنباط‌هایی در مورد توزیع‌های پسین مانند توزیع‌های حاشیه‌ای، میانگین و واریانس پارامترها و همچنین توزیع پیش‌بین مشاهدات آینده می‌باشند، فهم و بکارگیری این توزیع‌ها پایه‌ی استنباط بیزی را تشکیل می‌دهد. به‌طور مختصر اغلب مسائل استنباط آماری می‌تواند به صورت امید تابع مدنظر نسبت به توزیع پسین بیان شود. بنابراین با در دسترس بودن توزیع پسین، برای انجام استنباط‌های پسین از انتگرال‌گیری روی توزیع پسین استفاده می‌شود. از این رو توانایی انتگرال‌گیری توابع چند بعدی و پیچیده بسیار مهم می‌باشد و برای انجام استنباط بیزی نیازمند روش‌هایی هستیم که بتوانیم آن‌ها را در مسائل نمونه‌گیری از توزیع‌هایی با بعد بالا برای حل انتگرال‌ها به کار ببریم. یکی از این روش‌ها که معمولا بین آماردانان رایج است روش‌های مونت کارلو است که قصد داریم در این مقاله به آن بپردازیم. اما برای تقریب زدن انتگرال‌ها با استفاده از این روش، لازم است که نمونه هایی از توزیع پسین تولید شوند. در برخی موارد، صرفنظر از ثابت استانداردکننده، توزیع پسین معلوم است که در بخش دوم مقاله روش‌های مناسب برای تولید نمونه در این موارد بررسی خواهند شد. در مواردی نیز نمی‌توان به طور مستقیم از توزیع پسین نمونه‌گیری کرد و در واقع توزیع پسین ناشناخته است. در این حالت، خلاصه‌های پسین برای تقریب شکل توزیع پسین به کار می‌روند که در بخش سوم، از روش زنجیر مارکف با استفاده از مونت کارلو به عنوان یک راه‌حل مناسب برای تولید نمونه در چنین حالتی استفاده می شود.

واژه‌های کلیدی: قضیه بیز، تولید نمونه تصادفی، مونت کارلو، زنجیر مارکف با استفاده از مونت کارلو.

۱ مقدمه

فرض کنید مشاهدات y از یک چگالی نمونه‌ای $f(\mathrm{y}|\boldsymbol{\theta})$ به دست آمده باشند و بردار پارامتر $\boldsymbol{\theta}$ دارای توزیع پیشین $g(\boldsymbol{\theta})$ باشد، آنگاه طبق قضیه‌ی بیز چگالی پسین $\boldsymbol{\theta}$ متناسب خواهد بود با:

$$g(\boldsymbol{\theta}|\mathrm{y}) \propto g(\boldsymbol{\theta}) f(\mathrm{y}|\boldsymbol{\theta}).$$

مسئله ی موردنظر ما، خلاصه کردن این توزیع احتمال چند متغیره به منظور استنباط توابعی از $\boldsymbol{\theta}$ است. برای مثال فرض کنید علاقه‌مند به محاسبه‌ی میانگین پسین تابع

$h(\boldsymbol{\theta})$ هستیم. این مقدار میانگین را می‌توان به صورت نسبتی از انتگرال‌ها و به فرم زیر محاسبه نمود:

$$E(h(\boldsymbol{\theta})|\mathrm{y}) = \frac{\int h(\boldsymbol{\theta})g(\boldsymbol{\theta})f(\mathrm{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int g(\boldsymbol{\theta})f(\mathrm{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}.$$

همچنین برای محاسبه‌ی احتمال پسین قرار گرفتن $h(\boldsymbol{\theta})$ در مجموعه‌ای مثل A، کمیت زیر محاسبه می‌گردد:

$$p(h(\boldsymbol{\theta}) \in A|\mathrm{y}) = \frac{\int_{h(\boldsymbol{\theta})\in A} g(\boldsymbol{\theta})f(\mathrm{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int g(\boldsymbol{\theta})f(\mathrm{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}.$$

اگر بخواهیم توزیع حاشیه‌ای یک پارامتر را محاسبه کنیم، از توزیع پسین نسبت به بقیه پارامترها انتگرال گرفته می‌شود. بنابراین عملگر انتگرال نقشی اساسی در استنباط بیزی ایفا می‌کند. برای اغلب مدل‌های غیرخطی و مدل‌های خطی با بعد بالای پارامتر، اجرای قضیه‌ی بیز سخت می‌شود زیرا از حل برخی انتگرال‌ها ناتوان هستیم و یا اینکه حل آن‌ها بسیار پیچیده است و یک ارزیابی صریح از این انتگرال‌ها در دسترس نیست. از میان روش‌های تقریبی متفاوت می‌توان روش‌های مونت‌کارلو را نام برد که قصد داریم در این مقاله به آن بپردازیم.

۲ محاسبه‌ی انتگرال‌ها به روش مونت کارلو

روش‌های مونت کارلو از تکنیک‌های نمونه‌گیری تصادفی براساس شبیه‌سازی کامپیوتری برای تقریب انتگرال‌ها استفاده می‌کنند. ایده‌ی انتگرال‌گیری مونت کارلو را انریکو فرمی[۱] [۱] و استانیسلو اولام[۲] [۷] ارائه دادند.

[۱] Enrico Fermi
[۲] Stanislaw Ulam

در واقع علاقه‌مند به ارزیابی انتگرال‌ها از لحاظ عددی می‌باشیم، برای این منظور برخی روش‌ها وجود دارند که معمولا محاسبه‌ی انتگرال را پیچیده می‌کنند. هدف ما استفاده از روش‌هایی است که برای مسائل بیزی با بعد بالا نیزکارا باشند. در حالتی که توزیع پسین فرم تابعی استانداردی نداشته باشد، برای به دست آوردن استنباط‌های پسین می‌توان از یک تقریب ساده برای توزیع پسین استفاده کرد و با بکارگیری خلاصه‌های پسین یک شکل[۳] تقریبی از توزیع پسین را به دست آورد. یک روش کلی برای تقریب و خلاصه‌سازی توزیع پسین، براساس رفتار توزیع پسین حول نمای آن می‌باشد. در این روش، یک تقریب نرمال چندمتغیره برای پسین در نظر گرفته می‌شود که به عنوان یک تقریب اولیه‌ی خوب بوده و در آن بر پایه‌ی تقریب لاپلاس[۴] عمل می‌شود [۶]. روش دوم براساس شبیه‌سازی عمل کرده و از روش‌های $MCMC$ [۵] برای شبیه‌سازی نمونه‌هایی از پسین استفاده می‌کند که در ادامه به این روش خواهیم پرداخت.

۱.۲ روش برآورد مونت کارلو

روشی که در اینجا برای خلاصه‌سازی توزیع پسین به آن پرداخته می‌شود، استفاده از روش‌های محاسباتی بیزی بر اساس شبیه‌سازی می‌باشد. فرض کنید هدف ما محاسبه‌ی انتگرال زیر باشد،

$$\theta = \int_a^b g(x)\mathrm{d}x.$$

[۳] Shape
[۴] Laplace Approximation
[۵] Marcove Chain Monte Carlo

برای این منظور توزیع دلخواهی روی فاصله‌ی $[a,b]$ با چگالی $f(x)$ انتخاب می‌کنیم. متغیر تصادفی x را در فاصله $[a,b]$ و با چگالی $f(x)$ و متغیر تصادفی $h(x)$ را به صورت زیر تعریف می کنیم،

$$h(x) = \frac{g(x)}{f(x)}.$$

در این صورت خواهیم داشت،

$$\begin{aligned} \theta &= E(h(X)) \\ &= \int_a^b \frac{g(x)}{f(x)} f(x)\mathrm{d}x. \end{aligned}$$

اگر از چگالی f نمونه تصادفی x_1، x_2،...،x_n را استخراج کنیم، آنگاه $h(x_1)$،...، $h(x_n)$ نیز یک نمونه تصادفی است و داریم،

$$E(\frac{1}{n}\sum_{i=1}^{n} h(X_i)) = \theta.$$

طبق قانون قوی اعداد بزرگ[6]، با احتمال یک رابطه‌ی زیر برقرار است،

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \xrightarrow{a.s.} \theta.$$

بنابراین $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ برآورد معقولی برای θ می باشد. این برآورد، معادل با تقریب انتگرال مورد نظر است. با این روش، بسیاری از انتگرال‌هایی را که نمی توان با روش تحلیلی حل کرد، تقریب می‌زنند.

مثال ۱.۲. *می‌خواهیم انتگرال تابع زیر را روی فاصله‌ی* $[0,1]$ *محاسبه کنیم،*

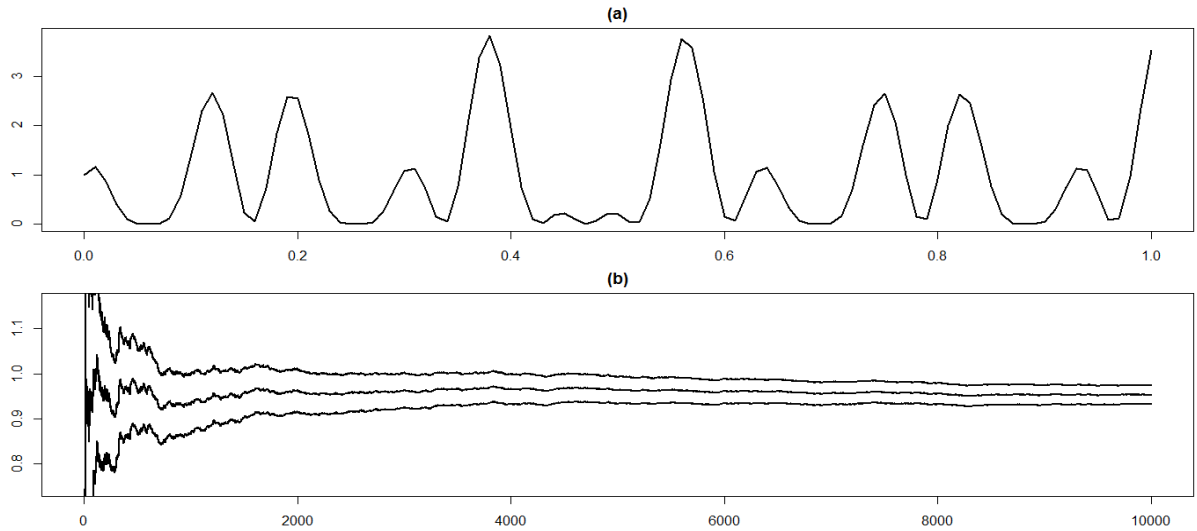
$$h(x) = [cos(50x) + sin(20x)]^2.$$

برای محاسبه‌ی این انتگرال، ابتدا متغیرهای تصادفی مستقل و هم‌توزیع با توزیع $U(0,1)$ *را تولید نموده و سپس کمیت* $\int h(x)\mathrm{d}x$ *را با استفاده از* $\sum h(U_i)/n$ *تقریب می‌زنیم. در شکل ۱، نمودار* (a) *تابع* $h(x)$ *را نشان می‌دهد و نمودار* (b)*، نشان‌دهنده‌ی میانگین‌ها و فواصل اطمینان به دست آمده از خطاهای استاندارد در مقابل تعداد شبیه‌سازی* n *است. همان‌طور که مشاهده می‌کنید با افزایش تعداد شبیه‌سازی، مقدار تقریبی به مقدار واقعی همگرا می‌شود. (برای دسترسی به کد مربوطه در نرم افزار* R *به نویسنده مراجعه شود.)*

۲.۲ استخراج نمونه از چگالی f

روش مونت کارلو یک روش موثر برای خلاصه‌سازی توزیع پسین است، در صورتی که بتوانیم نمونه‌های شبیه‌سازی شده‌ای از توزیع پسین دقیق داشته باشیم. این روش از تکنیک‌های نمونه‌گیری تصادفی استفاده کرده و بر پایه‌ی تولید متغیرهای تصادفی از توزیع پسین استوار است. در حالتی که توزیع‌های پسین، فرم تابعی شناخته شده و معلومی داشته باشند، نمونه‌های شبیه‌سازی‌شده را می‌توان به طور مستقیم از این توزیع‌ها به دست آورد و در نتیجه برآوردهای مونت کارلو از میانگین پسین برای هر تابعی از پارامتر امکان‌پذیر است. اما در بسیاری از حالات، توزیع پسین فرم معلومی ندارد و ما نیازمند این هستیم که الگوریتم جایگزینی را برای تولید نمونه ی شبیه‌سازی شده داشته باشیم. متغیر تصادفی X با چگالی fرا در نظر بگیرید، دو روش مهم برای استخراج نمونه از این چگالی

[6] Strong Law of Large Numbers



شکل ۱: تقریب انتگرال تابع $h(x) = [cos(۵۰x) + sin(۲۰x)]^۲$.

روش **تبدیل معکوس تابع توزیع**[۷] و روش **نمونه‌گیری پذیرش / رد**[۸] می‌باشد.

۱.۲.۲ روش تبدیل معکوس تابع توزیع

برای متغیر تصادفی X با چگالی f داریم،

$$F(X) \sim U(۰, ۱).$$

حال اگر از توزیع $U(۰, ۱)$ نمونه تصادفی $U_۱, ..., U_n$ استخراج شود، آنگاه با تبدیل زیر $x_۱, ..., x_n$ یک نمونه تصادفی از چگالی f می‌باشد [۲].

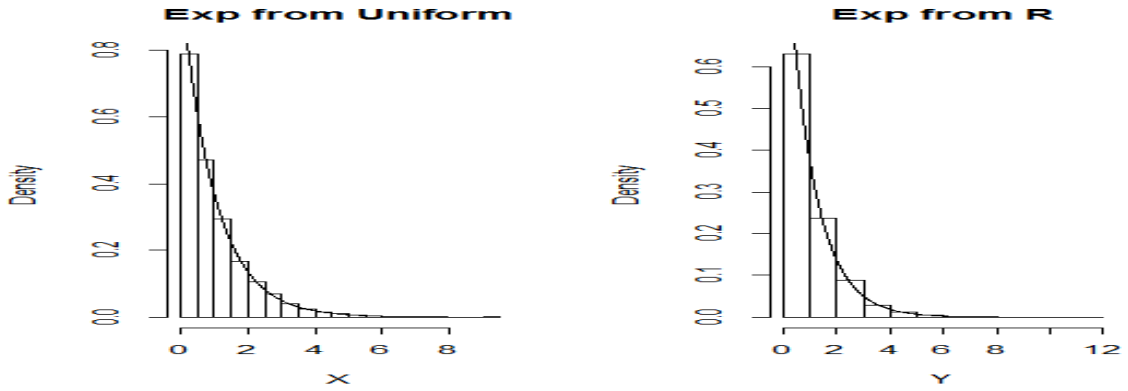
$$\begin{aligned} x_i &= F^{-۱}(U_i) \\ &= inf\{x : F(x) \geq U_i\}. \end{aligned}$$

مثال ۲.۲. *اگر X دارای توزیع نمایی با پارامتر $\lambda = ۱$ باشد، آنگاه تابع توزیع به فرم $F(x) = ۱ - e^{-x}$ خواهد بود. برای تولید مشاهدات از این توزیع، با استفاده از روش تبدیل انتگرال احتمال، با قرار دادن $u = ۱ - e^{-x}$ داریم؛ $x = -\log(۱ - u)$. بنابراین اگر $U \sim U(۰, ۱)$ باشد، آنگاه $X = -\log U \sim \exp(۱)$ خواهد شد. در شکل ۲ نمودار مرتبط با این توزیع را مشاهده می‌کنید. واضح است که نمودار حاصل از دو روش تبدیل معکوس و روش مستقیم، مشابه هستند. (برای دسترسی به کد مربوطه در نرم افزار R به نویسنده مراجعه شود.)*

۲.۲.۲ روش نمونه‌گیری پذیرش/ رد

در مواردی که توزیع معکوس‌پذیر نیست و یا به دست آوردن معکوس تابع پیچیده باشد، روش تبدیل معکوس عملکرد خوبی ندارد و نیازمند روش‌های دیگری هستیم. یکی از الگوریتم‌های مناسب برای شبیه‌سازی برآمدهای تصادفی از توزیع احتمال داده شده، نمونه‌گیری پذیرش/رد است. فرض کنید بخواهیم نمونه‌ای مستقل از چگالی

[۷] Inverse Transform Method
[۸] Accept-Reject Algorithm



شکل ۲: نمودار تولید متغیرهای تصادفی نمایی با استفاده از تبدیل معکوس (سمت راست) و با استفاده از دستور $rexp$ (روش مستقیم) در R (سمت چپ) مشاهده می‌شود. منحنی چگالی $\exp(۱)$ نیز در بالای این نمودار رسم شده است.

پسین $g(\theta|\mathrm{y})$ را که ثابت استانداردکننده‌ی آن مجهول است، داشته باشیم. اولین گام در نمونه‌گیری پذیرش/رد، یافتن چگالی احتمال دیگری مثل $p(\theta)$ می‌باشد. این چگالی پیشنهادی باید به گونه‌ای باشد که نمونه‌هایی از چگالی هدف $g(\theta|\mathrm{y})$ را فراهم کند و دارای ویژگی های زیر باشد:

۱. شبیه‌سازی برآمدها از $p(\theta)$ آسان باشد.

۲. چگالی p بر حسب مکان[۹] و وسعت[۱۰] مشابه چگالی پسین g مورد نظر باشد.

۳. برای θها و ثابت c داشته باشیم،

$$g(\theta|\mathrm{y}) \leq cp(\theta).$$

حال فرض کنید که چگالی p با این خصوصیات را یافته‌ایم. پس می‌توانیم با استفاده از الگوریتم پذیرش/رد به صورت زیر، برآمدهایی از چگالی پسین را به دست آوریم،

۱. به‌طور مستقل، θ از چگالی p و متغیر تصادفی یکنواخت U روی فاصله ی واحد شبیه‌سازی می‌شود.

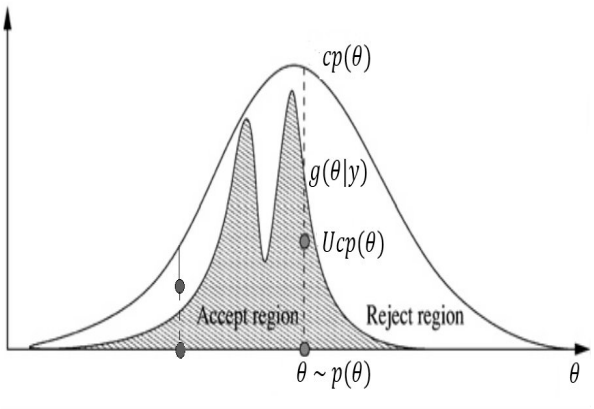
۲. اگر $U \leq \frac{g(\theta|\mathrm{y})}{cp(\theta)}$ باشد، آنگاه θ را به‌عنوان برآمدی از چگالی g پذیرفته، در غیر اینصورت θ رد می‌شود.

۳. تا زمانی که تعداد کافی از θهای مورد پذیرش فراهم شود، گام‌های ۱ و ۲ تکرار می‌شوند.

شکل ۳ نحوه‌ی اجرای این الگوریتم را نشان می‌دهد. در این شکل، چگالی هدف، چگالی پیشنهادی و نواحی رد و قبول را مشاهده می‌کنید. همان‌طور که می‌بینید برخی θهای انتخابی از چگالی پیشنهادی در ناحیه قبول و برخی دیگر در ناحیه رد قرار می‌گیرند.

یکی از وظایف اصلی در اجرای الگوریتم نمونه‌گیری رد، یافتن چگالی پیشنهادی p و ثابت c است. در گام ۲ الگوریتم، احتمال قبول برآمد کاندید شده به صورت $g(\theta|\mathrm{y})/cp(\theta)$ داده شده است. با محاسبه‌ی نسبت برآمدهایی از چگالی p که قبول شده‌اند، می‌توان در مورد کارایی الگوریتم اظهار نظر کرد. پس هر چه نرخ

[۹] Location
[۱۰] Spread



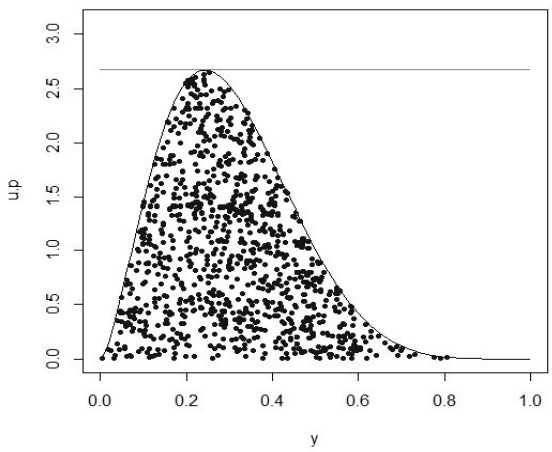
شکل ۳: الگوریتم پذیرش/رد: انتخاب θ از توزیع پیشنهادی و انتخاب یک متغیر تصادفی U از توزیع یکنواخت. قبول نمونه‌ی کاندیدی اگر $U \leq \frac{g(\theta|y)}{cp(\theta)}$ باشد و در غیر اینصورت رد آن.

قبول بالاتر باشد، الگوریتم کاراتر است. همچنین احتمال پذیرش θی استخراج شده از چگالی p، $1/c$ می‌باشد و این نشان می‌دهد که اگر c بزرگ باشد، نرخ پذیرش $1/c$ کوچک بوده و الگوریتم عملکرد خوبی ندارد. موضوع مهم این است که چگونه می‌توان c را کوچک ساخت؟

بدیهی است که مقدار c به چگالی p انتخاب شده بستگی دارد. اگر چگالی پیشنهادی p شبیه چگالی g باشد، آنگاه به مقدار c بزرگ احتیاجی نخواهد بود. پس چگالی p-ای انتخاب می‌کنیم که به‌سادگی قابل نمونه‌گیری بوده و هر چه بیشتر مشابه چگالی g باشد. البته، اگر چگالی g دم سنگین‌تر از چگالی p باشد، آنگاه c مناسبی نمی‌توان پیدا کرد. یکی از محدودیت‌های روش پذیرش/رد این است که یافتن تابع چگالی p-ای که به سادگی قابل نمونه‌گیری باشد و نسبت g/p کراندار بماند، همیشه کار چندان ساده‌ای نیست. به علاوه ممکن است که چنین چگالی p-ای را بیابیم ولی c، عدد بزرگی شود و در نتیجه احتمال پذیرش $1/c$، ناچیز بوده و این، کارایی روش پذیرش/رد را به شدت کاهش می‌دهد.

مثال ۳.۲. *می‌خواهیم با استفاده از روش پذیرش/رد و توزیع پیشنهادی یکنواخت، مشاهداتی از توزیع بتا با α و βی بزرگتر از یک تولید کنیم.*

بنابراین، کران بالای c، مقدار ماکزیمم چگالی بتا است و چون چگالی پیشنهادی p برابر با یک است، مقدار پیشنهادشده‌ی y پذیرفته می‌شود اگر $c \times U < g(y)$ باشد. توجه کنید که تولید $U \sim U(0,1)$ و ضرب آن در c معادل با تولید $U \sim U(0,c)$ می‌باشد.



شکل ۴: تولید متغیرهای تصادفی $x \sim \beta e(2.7, 6.3)$ با استفاده از الگوریتم پذیرش/رد

ابتدا ۲۵۰۰ جفت (y,U) به ترتیب از p و $U(0,c)$ تولید می‌شود. در شکل ۴ خط افقی نشان‌دهنده‌ی کران بالای c می‌باشد که برابر با ۲/۶۷ است. نقاط سیاه (y, Up) که زیر چگالی g قرار گرفته‌اند، نقاطی هستند که به عنوان مشاهداتی از چگالی g پذیرفته شده و نقاطی که خارج از این ناحیه قرار می‌گیرند، رد شده‌اند. در واقع به طور تقریبی $1/2.67 = 37$ درصد از مقادیر پذیرفته شده‌اند.(برای دسترسی به کد مربوطه در نرم افزار R به نویسنده مراجعه شود.)

بنابراین به طور خلاصه در انتگرال‌گیری به روش مونت‌کارلو، با استفاده از تعداد زیادی شبیه‌سازی و تقریب $\frac{1}{M}\sum_{i=1}^{M} g(\theta^i)$، مقدار انتگرال $\int_\theta g(\theta)p(\theta)\mathrm{d}\theta$ را برآورد می‌کنیم، که با توجه به قضیه ی $SLLN$، این تقریب برآورد سازگاری را برای انتگرال مدنظر، به ما می‌دهد.
سوالی که در اینجا مطرح می‌باشد این است که اگر نتوانیم نمونه‌های مستقل تولید کنیم، چه خواهد شد؟
در ادامه‌ی مقاله به این موضوع می‌پردازیم.

۳ کاربرد زنجیرهای مارکف با استفاده از مونت کارلو $(MCMC)$ در مسائل بیزی

زنجیر مارکف با استفاده از مونت کارلو $(MCMC)$، روش جایگزینی را فراهم می‌کند که به وسیله‌ی آن می‌توان از توزیع پسین به‌طور مستقیم نمونه‌گیری کرده و انتگرال‌گیری به‌طور صریح انجام شود.
ایده‌ی نمونه‌گیری $MCMC$ را اولین بار متروپلیس[۱۱] و همکارانش [۵] در سال ۱۹۵۳، به‌عنوان روشی برای شبیه‌سازی از سطوح انرژی اتم‌ها در یک ساختار بلورین معرفی کردند و بعدها در سال ۱۹۷۰ توسط هستینگ[۱۲] [۴] بهبود یافته و به مسائل آماری تعمیم داده شد.

[۱۱] Metropolis
[۱۲] Hastings

دیدیم که برای تقریب کمیت،

$$\begin{aligned}\theta &= E(h(X))\\ &= \int_a^b g(x)\mathrm{d}x\\ &= \int_a^b \frac{g(x)}{f(x)} f(x)\mathrm{d}x,\end{aligned}$$

به روش مونت کارلو، با استخراج مشاهدات مستقل و هم‌توزیع، از چگالی $f(x)$و برآورد زیر استفاده می کنیم،

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

اما استخراج نمونه از چگالی $f(x)$ مسئله‌ی اساسی روش مونت کارلو می‌باشد. در واقع به دنبال روش قدرتمندتری برای استخراج نمونه هستیم که نسبت به روش‌های قبل محدودیت‌های کمتری داشته باشد. برای مثال، فرض کنید با توجه به نبود ثابت استانداردکننده، برآمدهایی که از توزیع پسین (توزیع هدف) به دست می آیند مستقل نباشند، بنابراین می‌توان با استفاده از زنجیر مارکف برآمدهایی با وابستگی کم را تولید نمود و سپس خلاصه‌های پسین را با استفاده از این برآمدها به دست آورد. صرفنظر از اینکه نقطه شروع الگوریتم چه باشد، برای همه ی الگوریتم‌های $MCMC$، توزیع زنجیره مارکف تولید شده به توزیع هدف همگراست. بنابراین اگر بتوانیم زنجیره مارکفی را در نظر بگیریم که توزیع ایستای[۱۳] آن، توزیع پسین $g(\theta|y)$ موردنظر باشد، آنگاه می‌توانیم این زنجیره را برای به دست آوردن برآمدهایی از $g(\theta|y)$ اجرا کنیم، تا زمانی که زنجیره همگرا شود. چون برآمدهای این زنجیره مارکف وابستگی دارند، نمی‌توان از قضیه ی $SLLN$ برای همگرایی استفاده کرد ولی خوشبختانه با

[۱۳] Stationary Distribution

توجه به قضیه ی ارگودیکی[۱۴] و شرایط آن [۳] می‌توان از وابستگی بین برآمدهای زنجیره مارکف صرفنظر کرد و کمیت‌های موردنظر در استنباط آماری را محاسبه نمود. در این موارد یک فرآیند نمونه‌گیری مونت کارلو تکراری را اتخاذ کرده، که به دلیل همگرایی، برآمدی تصادفی از توزیع پسین توام موردنظر را تضمین می‌کند. این فرایندهای مونت کارلو تکراری، به طور خاص، دنباله‌ای تصادفی با ویژگی مارکفی(ارگودیکی) را تولید می‌کنند. حال مسئله این است که چگونه می‌توان چنین زنجیره‌ای را ایجاد نمود. یکی از راه‌های معمول برای شبیه‌سازی از توزیع پسین، استفاده از روش‌های مونت‌کارلو با زنجیره مارکف ($MCMC$) می‌باشد. روش نمونه‌گیری $MCMC$، زنجیره مارکفی نامتناوب و تحویل‌ناپذیر را برای توزیع ایستایی برابر با توزیع پسین موردنظر، ایجاد می‌کند. دو الگوریتم مهم برای تولید چنین زنجیره‌ای، الگوریتم‌های **متروپلیس - هستینگ** [۱۵] و **نمونه‌گیر گیبز** [۱۶] می‌باشند. در ادامه این دو الگوریتم را مورد بررسی قرار خواهیم داد.

۱.۳ الگوریتم متروپلیس - هستینگ

یک راه کلی برای ایجاد زنجیره مارکف، استفاده از الگوریتم متروپلیس - هستینگ می‌باشد.
فرض کنید بخواهیم از چگالی پسین $g(\theta|y)$ شبیه‌سازی کنیم، که برای راحتی این چگالی را با نماد $g(\theta)$ نشان می‌دهیم. الگوریتم متروپلیس - هستینگ با مقدار اولیه ی θ° و تعیین قاعده‌ای برای $\theta^{(t-1)}$، شروع می‌شود. این قاعده، شامل یک چگالی پیشنهادی است که مقدار کاندید θ^* را شبیه‌سازی می‌کند و هم‌چنین شامل احتمال قبول P می‌باشد که بیانگر احتمال قبول مقدار کاندید به عنوان مقدار بعدی دنباله است.
این الگوریتم را می‌توان به صورت زیر شرح داد:

۱. مقدار کاندید θ^* را از چگالی پیشنهادی $p(\theta^*|\theta^{(t-1)})$ شبیه‌سازی کنید.

۲. نسبت $R = \frac{g(\theta^*)p(\theta^{(t-1)}|\theta^*)}{g(\theta^{(t-1)})p(\theta^*|\theta^{(t-1)})}$ را محاسبه کنید.

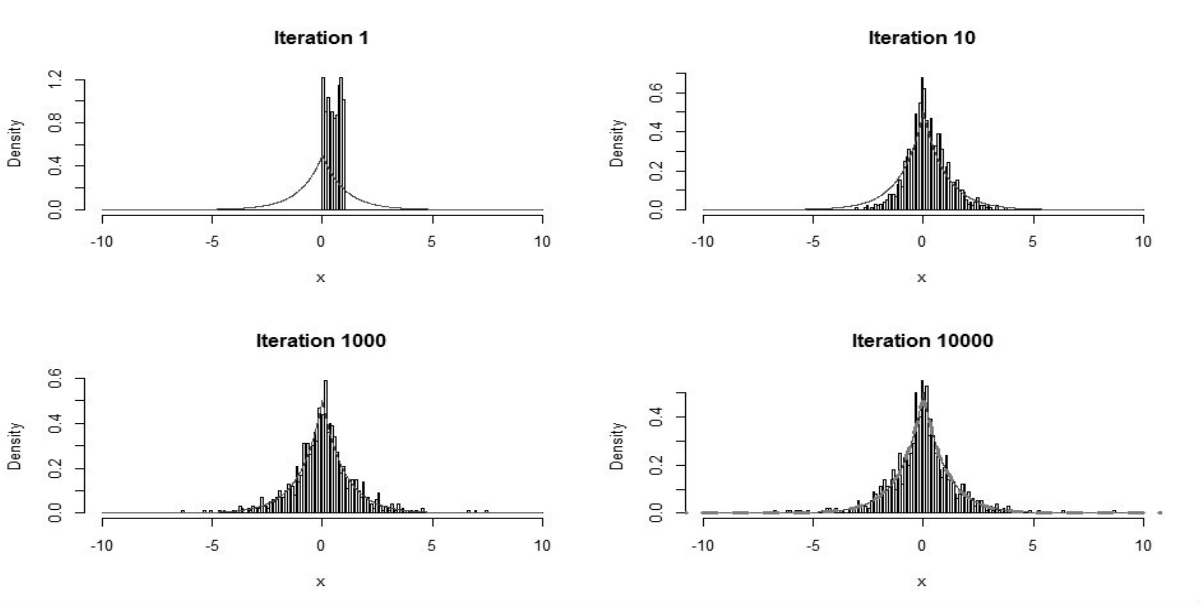
۳. احتمال قبول $P = min\{R, 1\}$ را محاسبه کنید.

۴. مقدار $\theta^{(t)}$ را به صورت زیر انتخاب کنید.
- با احتمال P؛ $\theta^{(t)} = \theta^*$،
- با احتمال $1 - P$؛ $\theta^{(t)} = \theta^{(t-1)}$.

تحت بعضی شرایط نظم روی چگالی پیشنهادی $p(\theta^*|\theta^{(t-1)})$، دنباله برآمدهای شبیه‌سازی شده ی θ^1، θ^2، ... به متغیری همگراست که با توجه به چگالی پسین $g(\theta)$ توزیع شده است [۸].

مثال ۱.۳. *می‌خواهیم دنباله‌ی x_1، ...، x_n را طوری شبیه‌سازی کنیم که هر x_i دارای توزیع نمایی دوگانه و با چگالی $f(x) = \alpha/2exp(-\alpha|x|)$ باشد. فرض کنید فقط می‌دانیم که چگونه از متغیرهای تصادفی $N(0, 1)$ شبیه‌سازی کنیم. بنابراین چگالی پیشنهادی، چگالی $N(0, 1)$ روی $\mathcal{R}$ می‌باشد. حال با استفاده از این چگالی پیشنهادی و با اجرای گام‌های الگوریتم $M-H$ نمونه‌هایی را از چگالی هدف(نمایی دوگانه) به‌دست می‌آوریم. در شکل ۵ نمودارهای مرتبط با چهار تکرار مختلف الگوریتم مشاهده می‌شود. با افزایش تعداد تکرارها، برآمدهای الگوریتم به چگالی نمایی دوگانه نزدیکتر می‌شوند. (برای دسترسی به کد مربوطه در نرم افزار R به نویسنده مراجعه شود.)*

[۱۴] Ergodic Theorem
[۱۵] Metropolis-Hastings
[۱۶] Gibbs Sampler



شکل ۵: نمودار مربوط به اجرای الگوریتم متروپلیس - هستینگ برای شبیه‌سازی نمونه‌هایی از چگالی نمایی دوگانه.

۲.۳ نمونه‌گیری گیبز

یکی دیگر از روش‌ها برای تولید زنجیر مارکف، الگوریتم نمونه‌گیر گیبز است که حالت خاصی از روش متروپلیس هستینگ می‌باشد. فرض کنید بردار پارامترهای مورد نظر به صورت $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ باشد. توزیع توام پسین $\boldsymbol{\theta}$، که با $[\boldsymbol{\theta}|data]$ نشان داده می‌شود، ممکن است بعد بالایی داشته و برای خلاصه‌سازی سخت باشد.

فرض کنید مجموعه‌ای از توزیع‌های شرطی به صورت زیر را داریم،

$$[\theta_1|\theta_2, ..., \theta_p, data],$$

$$[\theta_2|\theta_1, \theta_3, ..., \theta_p, data],$$

$$\vdots$$

$$[\theta_p|\theta_1, ..., \theta_{p-1}, data].$$

$[x|Y, Z]$ بیانگر توزیع شرطی x روی مقادیری از متغیرهای تصادفی Y و Z است. ایده ی نمونه‌گیر گیبز این است که می‌تواند الگوریتم شبیه‌سازی از توزیع پسین را با شبیه‌سازی پارامترهای تک از مجموعه توزیع‌های شرطی تنظیم کند. شبیه‌سازی برآمدی از هر پارامتر تک با این توزیع‌ها، حلقه‌ای از نمونه‌گیر گیبز نامیده می‌شود. تحت شرایط کلی، برآمدهای این الگوریتم شبیه‌سازی به توزیع هدف (توزیع پسین توام $\boldsymbol{\theta}$) همگراست [۸]. پس اگر توزیع شرطی هر پارامتر به شرط اطلاعات معلوم و دیگر پارامترها، یعنی $p(\theta_j|\theta_{(-j)}, \mathrm{y})$ را بدانیم، می‌توان از آن‌ها برای نمونه‌گیری از توزیع توام (برای مثال توزیع پسین) استفاده کرد. سوالی که ممکن است در ذهن پیش بیاید این است که چگونه می‌توان توزیع توام را با تشخیص توزیع‌های شرطی تعیین کرد؟ برای پاسخ به این سوال قضیه‌ی بعد مطرح می‌شود.

قضیه ۱.۳. *قضیه برای دو بلوک بیان می‌گردد.*

فرض کنید چگالی پسین $f(\mathrm{x},\mathrm{y})$ را داشته باشید. چگالی پسین را می‌توان بر حسب چگالی‌های شرطی $f(\mathrm{x}|\mathrm{y})$ و $f(\mathrm{y}|\mathrm{x})$ به صورت زیر نوشت،

$$f(\mathrm{x},\mathrm{y}) = \frac{f(\mathrm{y}|\mathrm{x})}{\int \frac{f(\mathrm{y}|\mathrm{x})}{f(\mathrm{x}|\mathrm{y})}\mathrm{dy}}.$$

این قضیه را، قضیه‌ی هامرسلی-کلیفورد[۱۷] می‌نامند.

برهان. برای اثبات مخرج را می‌توان به صورت زیر نوشت،

$$\begin{aligned}\int \frac{f(\mathrm{y}|\mathrm{x})}{f(\mathrm{x}|\mathrm{y})}\mathrm{dy} &= \int \frac{\frac{f(\mathrm{x},\mathrm{y})}{f(\mathrm{x})}}{\frac{f(x,\mathrm{y})}{f(\mathrm{y})}}\mathrm{dy} \\ &= \int \frac{f(\mathrm{y})}{f(\mathrm{x})}\mathrm{dy} \\ &= \frac{۱}{f(\mathrm{x})}.\end{aligned}$$

بنابراین سمت راست تساوی اول به صورت زیر خواهد شد،

$$\begin{aligned}\frac{f(\mathrm{y}|\mathrm{x})}{\frac{۱}{f(\mathrm{x})}} &= f(\mathrm{y}|\mathrm{x})f(\mathrm{x}) \\ &= f(\mathrm{x},\mathrm{y}).\end{aligned}$$

قضیه بیان می‌کند که اطلاعات در مورد چگالی‌های شرطی به ما اجازه‌ی دستیابی به چگالی توام را می‌دهد که برای بیش از دو بلوک از پارامترها نیز کاربرد دارد. حال می‌خواهیم بررسی کنیم که چطور می‌توان چگالی‌های شرطی را به دست آورد. فرض کنید پسین $g(\boldsymbol{\theta}|\mathrm{y})$ را داشته باشیم، برای محاسبه‌ی توزیع‌های شرطی برای هر پارامتر، به صورت زیر عمل می‌کنیم؛

۱. چگالی پسین را صرفنظر از ثابت‌های تناسب می‌نویسیم.

۲. برای مثال پارامتر $\theta_۱$ را انتخاب و هر عبارتی که به $\theta_۱$ بستگی نداشته باشد را کنار می‌گذاریم.

۳. با استفاده از اطلاعات و تجربه‌ی شخصی، تشخیص دهیم که ثابت استانداردکننده چیست و در نتیجه توزیع شرطی کامل $g(\theta_۱|\theta_{(-۱)},\mathrm{y})$، تعیین می‌شود.

۴. گام‌های ۲ و ۳ را برای همه‌ی بلوک‌های پارامتر تکرار می‌کنیم.

۱.۲.۳ گام‌های نمونه‌گیر گیبز

فرض کنید علاقه‌مند به نمونه‌گیری از توزیع پسین $g(\boldsymbol{\theta}|\mathrm{y})$ هستیم که $\boldsymbol{\theta}$ برداری از سه پارامتر $\theta_۱$ و $\theta_۲$ و $\theta_۳$ است. گام‌های نمونه‌گیر گیبز به صورت زیر خواهند بود؛

۱. انتخاب برداری از مقادیر اولیه‌ی $(\theta^{(۰)})$ (تعریف توزیع اولیه‌ی شرطی(شروع) و انتخاب $\theta^{(۰)}$ از آن).

۲. شروع با هر θ دلخواه (در اینجا، ترتیب مهم نیست ولی برای راحتی با $\theta_۱$شروع می‌کنیم.) مقدار $\theta_۱^{(۱)}$ را از توزیع شرطی $g(\theta_۱|\theta_۲^{(۰)},\theta_۳^{(۰)},\mathrm{y})$ بیرون می‌کشیم.

۳. مقدار $\theta_۲^{(۱)}$ (باز هم ترتیب مهم نیست) را از توزیع شرطی کامل به فرم $g(\theta_۲|\theta_۱^{(۱)},\theta_۳^{(۰)},\mathrm{y})$ انتخاب می‌کنیم. توجه کنید که در اینجا مقدار به‌روز $\theta_۱^{(۱)}$ را استفاده می‌کنیم.

۴. مقدار $\theta_۳^{(۱)}$ را از توزیع شرطی کامل به فرم $g(\theta_۳|\theta_۱^{(۱)},\theta_۲^{(۱)},\mathrm{y})$ به دست می‌آوریم.

۵. $\boldsymbol{\theta}^{(۲)}$ را با استفاده از $\boldsymbol{\theta}^{(۱)}$ و به طور پیوسته با استفاده از به‌روزترین مقادیر، به دست می‌آوریم.

[۱۷] Hammersley-Clifford

۶. گام‌ها را تا زمانی تکرار می‌کنیم که M برآمد موردنظر به دست آید.

۴ بحث و نتیجه‌گیری

یکی از مهم‌ترین مباحث در آمار، استنباط آماری است. در مدل‌های بیزی نیز معمولا استنباط‌ها بر اساس توزیع پسین صورت می‌پذیرد که اغلب لازمه‌ی چنین استنباط‌هایی، حل انتگرال‌های پیچیده است. علاوه بر روش‌های تحلیلی برای حل این انتگرال‌ها، روش‌هایی تقریبی نیز وجود دارند که در ابعاد بالا کارایی بیشتری داشته و از لحاظ محاسباتی بهینه می‌باشند. با توجه به اینکه توزیع پسین فرم مشخصی داشته باشد و یا اینکه نمونه‌گیری از آن راحت باشد از هر کدام از روش‌های مونت کارلو یا الگوریتم‌های $MCMC$ مثل متروپلیس هستینگ یا گیبز استفاده می‌شود.

مراجع

[1] De Gregorio, A. G. (2005). Neutron physics in the early 1930s. Historical Studies in the Physical and Biological Sciences, 35, 293–340.

[2] Devroye, L. (1986). Non-uniform random variate generation. New York: Springer-Verlag. Chapter 2, Section 2, 28.

[3] Foguel, S. R. (1969). The ergodic theory of Markov processes. New York, Van Nostrand Reinhold Co, 21.

[4] Hastings, W. (1970). Monte carlo sampling methods using markov chains and their application. Biometrika, 57, 97–109.

[5] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equations of state calculations by fast computing machines. J. Chem. Phys., 21, 1087–1092.

[6] Nocedal, J. and Wright, S. J. (1999). Numerical optimization. Springer-Verlag.

[7] Richtmyer , D., Pasta, J. and Ulam, S. (1947). Statistical methods in neutron diffusion. LANL report LAMS-551. Prenite 23 oktobron 2011.

[8] Roberts G. O. and SmithA, F. M. (1994). Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. 49(2), 207–216.