استنباط زنجیرهای مارکف: مروری بر مقایسه الگو، برآورد بیزی و نرخ آنتروپی

معصومه اصلی بیگی [۱] ، حمید پزشک [۲]

چکیده:

زنجیرهای مارکف در زمینه‌های مختلف علمی برای تحلیل داده‌های دنباله‌ای به‌کار می‌روند. در این نوشته با استفاده از مقاله [۳] نشان می‌دهیم چگونه می‌توان زنجیرهای مارکف از مرتبه دلخواه k را برای داده‌های متناهی با استفاده از روش‌های بیزی در برآورد پارامترها و انتخاب مرتبه الگو مورد استفاده قرار داد. با به کارگیری روش‌هایی از نظریه اطلاع، آنتروپی نسبی و نرخ آنتروپی را محاسبه می‌کنیم. در نهایت با ارائه یک مثال، مراحل بالا را که مراحل مختلف استنباط یک زنجیر مارکف متناهی است به صورت شهودی بیان می‌کنیم.

واژه‌های کلیدی: استنباط بیزی، توزیع پیشین، توزیع پسین، الگوی مارکف، نرخ آنتروپی.

۱ مقدمه

در بسیاری از سیستم‌های طبیعی، داده‌ها از دنباله‌ای از حروف از یک الفبای متناهی تشکیل شده‌اند. برای مثال، تحلیل اطلاع دنباله در بیوپلیمر از این دسته است [۵]. استنباط آماری الگوهایی که در این نوع سیستم‌ها دیده می‌شود از دیرباز مورد توجه محققان بوده است. این الگوهادر کلاس الگوهای زنجیرهای مارکف قرار می‌گیرند [۲و۴]. در اغلب مطالعات انجام شده برای تجزیه و تحلیل این الگوها، مراحل استنباط آماری به بخش‌های زیر تقسیم می‌شود: ۱. برآورد نقطه‌ای پارامترهای الگو ۲. مقایسه الگو ۳. برآورد توابعی از پارامترهای الگو.

استنباط بیزی همه این مراحل را به هم مرتبط می‌کند [۱]. هر سه مرحله نامبرده در ارتباط با یکدیگر هستند و به کارگیری هر یک از آنها بدون توجه به سایر مراحل سبب به وجود آمدن تفسیرهای گمراه کننده می‌شود. ترکیب استنباط پارامترهای الگو، مقایسه الگو و برآورد نرخ آنتروپی [۲و۴] به عنوان سومین مرحله، سبب درک بهتر الگوهای زنجیر مارکف می‌شود و این مسئله در مورد منبع تولید داده‌هایی که خارج از کلاس الگوی زنجیر مارکف قرار می‌گیرد نیز درست است. با مقایسه الگو، ساختار منبع داده‌ها معین می‌شود و با برآورد نمودن نرخ آنتروپی توصیفی از تصادفی بودن منبع داده‌ها فراهم می‌شود. این

[۱] کارشناس ارشد آمار، بانک مرکزی
[۲] استاد گروه آمار دانشگاه تهران

نوشته که اقتباسی از [۳] است در صدد است نوع متفاوتی از استنباط آماری را دوره کند. این استنباط برای دنباله‌ای متناهی از پیشامدها مورد توجه محققین است.

۲ استنباط پارامترهای الگو

در مرحله اول از استنباط بیزی یک ارتباط سیستماتیک بین داده‌ها D، الگوی انتخابی M و بردار پارامترهای الگو θ معرفی می‌شود. موضوع جالب توجه در استنباط پارامترهای الگو، چگالی احتمال پسین $P(\theta|D,M)$ است که احتمال پارامترهای الگو به شرط داده‌های مشاهده شده و الگوی انتخابی است. برای یافتن پسین در ابتدا، چگالی توأم $P(\theta,D|M)$ داده‌ها و پارامترهای الگو به شرط دانستن اینکه الگوی M انتخاب شده است، مورد بررسی قرار می‌گیرد. چگالی پسین بنابر قضیه بیز به دست می‌آید:

$$P(\theta|D,M)=\frac{P(D|\theta,M)P(\theta|M)}{P(D|M)} \qquad (۱)$$

چگالی پیشین $P(\theta|M)$ توزیع روی پارامترهای الگو را تعیین می‌کند. درستنمایی $P(D|\theta,M)$ احتمال داده‌ها را به شرط پارامترهای الگو و الگوی داده شده شرح می‌دهد. در خاتمه سندیت [۳] $P(D|M)$ احتمال داده‌ها به شرط الگوی داده شده است. در بخش بعدی هر یک از این کمیت‌ها برای به دست آوردن عبارتی صریح برای چگالی پیشین شرح داده می‌شود.

۳ زنجیرهای مارکف

اولین قدم در استنباط مشخص کردن فرضیات الگو است. فرض می‌کنیم که یک مجموعه داده D از طول N، نقطه شروع استنباط را نشان دهد. D از نمادهای s_t که متعلق به یک الفبای متناهی A است، انتخاب می‌شود:

$$D=s_0 s_1 ... s_{N-1} \quad s_t \in A \qquad (۲)$$

نماد $s_t^{\leftarrow k}$ برای نشان دادن یک دنباله با طول k از حروف که در موقعیت t پایان یافته است معرفی می‌شود: مثلاً $s_4^{\leftarrow 2}=s_3 s_4$. کلاس الگوی زنجیر مارکف مرتبه k ام، یک حافظه متناهی و همگن در منبع داده‌ها متصور می‌شود و می‌تواند به صورت زیر نوشته شود:

$$p(D)=p(s_{k-1}^{\leftarrow k})\prod_{t=k-1}^{N-2}p(s_{t+1}|s_t^{\leftarrow k}) \qquad (۳)$$

پارامترهای زنجیر مارکف مرتبه k ام عبارتند از:

$$\theta_k=\{p(s|s^{\leftarrow k}) : s\in A, s^{\leftarrow k}\in A^k\} \qquad (۴)$$

برای هر کلمه $s^{\leftarrow k}$ باید شرط $\sum_{s\in A}p(s|s^{\leftarrow k})=1$ برقرار باشد.

۱.۳ درستنمایی

با داده‌های $D=s_0 s_1 ... s_{N-1}$، درستنمایی می‌تواند با استفاده از خاصیت مارکف معادلهٔ (۳) نوشته شود:

$$P(D\mid\theta_k,M_k)=\prod_{s\in A}\prod_{s^{\leftarrow k}\in A^k}[p(s|s^{\leftarrow k})]^{n(s^{\leftarrow k}s)} \qquad (۵)$$

[۳] Evidence

که در آن $n(s^{\leftarrow k}s)$ تعداد زمان‌هایی است که کلمه $s^{\leftarrow k}s$ در نمونهٔ D رخ داده است. برای استفاده‌های بعدی، نماد $n(s^{\leftarrow k}) = \sum_{s\in A} n(s^{\leftarrow k}s)$ برای تعداد زمان‌هایی که کلمه $s^{\leftarrow k}$ مشاهده شده است، معرفی می‌شود. توجه می‌کنیم که معادله (۵) روی دنباله آغازین $s_{k-1}^{\leftarrow k} = s_0 s_1 ... s_{k-1}$ شرطی شده است.

۲.۳ چگالی پیشین

چگالی پیشین $P(\theta_k|M_k)$ برای مشخص کردن فرضیات در مورد الگو، قبل از اینکه داده‌ها در دسترس باشند، استفاده می‌شود. فرم دقیق پیشین به وسیله مفروضات ما از ابرپارامترها $\alpha(s^{\leftarrow k}s)$ برای پیشین معین می‌شود. در زنجیر مارکف مرتبه k ام یک ابرپارامتر برای هر کلمه $s^{\leftarrow k}s$ به ازای الفبای تحت بررسی داده شده وجود دارد. پیشین مزدوج برای استنباط زنجیر مارکف حاصلضربی از توزیع‌های دیریکله است که در آن برای هر کلمه $s^{\leftarrow k}$ چگالی دیریکله‌ای در نظر گرفته شده است [۸]. بنابراین:

$$P(\theta_k|M_k) = \prod_{s^{\leftarrow k}\in A^k} \{\frac{\Gamma[\alpha(s^{\leftarrow k})]}{\prod_{s\in A}\Gamma[\alpha(s^{\leftarrow k}s)]} \times \delta(1 - \sum_{s\in A} p(s|s^{\leftarrow k})) \times \prod_{s\in A} p(s|s^{\leftarrow k})^{\alpha(s^{\leftarrow k}s)-1}\} \quad (۶)$$

که در آن $\alpha(s^{\leftarrow k}) = \sum_{s\in A}\alpha(s^{\leftarrow k}s)$ و تابع δ عامل نرمال ساز برای پارامترهای الگو است، یعنی به‌ازای $t = 0$، $\delta(t) = 1$ و به‌ازای $t \neq 0$، $\delta(t) = 0$ است. با داشتن این فرم تابعی حداقل دو راه برای تفسیر آنچه که پیشین درباره پارامترهای زنجیر مارکف می‌گوید، وجود دارد: ۱. روش آمار کلاسیک: در این روش میانگین و واریانس $p(s|s^{\leftarrow k})$ را نسبت به توزیع پیشین بدست می‌آوریم. ۲. روش آمار بیزی: توزیع حاشیه‌ای را برای پارامتر هر الگو بررسی می‌کنیم. برای توزیع دیریکله توزیع حاشیه‌ای برای هر پارامتر توزیع بتا خواهد بود.

فرض عمومی در استنباط الگو این است که بپذیریم همه پارامترها دارای توزیع پیشین یکنواخت هستند، یعنی: $\alpha(s^{\leftarrow k}s) = 1 \quad \forall s \in A, s^{\leftarrow k} \in A^k$

۳.۳ سندیت

با درستنمایی داده شده و پیشین انتخاب شده، مدرک $P(D|M_k)$ جمله نرمال کننده در قضیه بیز است.

$$P(D \mid M_k) = \int P(D|\theta_k, M_k)P(\theta_k|M_k)d\theta_k \quad (۷)$$

با به کارگیری معادله فوق در درستنمایی معادله (۵) و پیشین معادله (۶)، سندیت به دست می‌آید:

$$P(D \mid M_k) = \prod_{s^{\leftarrow k}\in A^k} \{\frac{\Gamma[\alpha(s^{\leftarrow k})]}{\prod_{s\in A}\Gamma[\alpha(s^{\leftarrow k}s)]} \times \frac{\prod_{s\in A}\Gamma[n(s^{\leftarrow k}s) + \alpha(s^{\leftarrow k}s)]}{\Gamma[n(s^{\leftarrow k}) + \alpha(s^{\leftarrow k})]}\} \quad (۸)$$

۴.۳ چگالی پسین

با استفاده از قضیهٔ بیز معادله (۱) و نتایج قبلی، توزیع پسین روی پارامترهای زنجیر مارکف مرتبه k بدست می آید:

$$P(\theta_k|D,M_k) = \prod_{s^{\leftarrow k}\in A^k}\{\frac{\Gamma[n(s^{\leftarrow k})+\alpha(s^{\leftarrow k})]}{\prod_{s\in A}\Gamma[n(s^{\leftarrow k}s)+\alpha(s^{\leftarrow k}s)]} \times\delta(1-\sum_{s\in A}p(s|s^{\leftarrow k})) \times\prod_{s\in A}p(s|s^{\leftarrow k})^{n(s^{\leftarrow k}s)+\alpha(s^{\leftarrow k}s)-1})\} \quad (۹)$$

چگالی پسین توزیع دیریکله با پارامترهای تغییر یافته نسبت به توزیع پیشین است که نتیجهٔ انتخاب پیشین مزدوج است. همانند چگالی پیشین، دو راه برای فهمیدن آنچه که چگالی پسین درباره تغییرات درپارامترهای زنجیر مارکف برآورد شده به ما می‌گوید، وجود دارد. اولی، استفاده از برآورد نقطه‌ای است و از میانگین و واریانس $p(s|s^{\leftarrow k})$ نسبت به پسین به دست می آید. دوم بررسی توزیع حاشیه‌ای برای پارامتر هر الگو است. برای توزیع دیریکله، توزیع حاشیه‌ای برای هر پارامتر توزیع بتا خواهد بود.

۴ مقایسه الگو

مجموعه الگوهای مرتبهk، $M=\{M_k\}_{k_{min}}^{k_{max}}$ را در نظر بگیرید. احتمال توأم $P(M_k,D|M)$ از یک الگوی بخصوص $M_k\in M$ و داده‌های D بنابر قضیه بیز به صورت زیر است:

$$P(M_k|D,M)=\frac{P(D|M_k,M)P(M_k|M)}{P(D|M)} \quad (۱۰)$$

که در آن مخرج یک مجموع است و بنابر قانون کل احتمال با مجموع گیری روی الگوهای $M_k\in M$ به دست می آید. احتمال یک الگوی ویژه در مجموعهٔ تحت بررسی با دو مؤلفه داده می‌شود: مدرک $P(D|M_k,M)$ که از معادله (۸) مشخص می‌شود و پیشین روی رتبه الگو $P(M_k|M)$.

دو پیشین متداول در مقایسهٔ الگو عبارتند از: ۱. پیشین یکنواخت که در آن همه الگوها به طور مساوی محتمل هستند. ۲. الگوهایی که باید برای تعداد پارامترهای آزاد استفاده شده برای برازش داده‌ها جریمه در نظر بگیرند. در اولین مورد، $P(M_k|M)=\frac{1}{|M|}$ برای همه مراتب k یکسان است. احتمال الگوهایی که از این پیشین استفاده می‌کنند، برابر است با:

$$P(M_k|D,M)=\frac{P(D|M_k,M)}{\sum_{M'_k\in M}P(D|M'_k,M)} \quad (۱۱)$$

در مورد دوم که یک جریمه برای تعداد پارامترهای الگو در نظر گرفته می‌شود با استفاده از فرضی که به الگوهای با پارامتر زیاد جریمه بیشتر تعلق می‌دهد، انتخاب الگو را بر اساس رابطه زیر انجام می‌دهیم:

$$P(M_k|D,M)=\frac{P(D|M_k,M)\exp(-|M_k|)}{\sum_{M'_k}P(D|M'_k,M)\exp(-|M'_k|)} \quad (۱۲)$$

که در آن $|M_k|$ تعداد پارامترهای آزاد الگو است. برای زنجیر مارکف مرتبه k ام، تعداد پارامترهای

آزاد برابر است با

$$| M_k | = |A|^k (|A| - ۱) \qquad (۱۳)$$

که در آن $|A|$ اندازه A است.

۵ نظریه اطلاع، مکانیک آماری و نرخ‌های آنتروپی

ارتباط بین استنباط و نظریه اطلاع [٦] با در نظر گرفتن معادله پیشین (٦) و معادله درستنمایی (۵) آغاز می‌شود.

$$P(\theta_k|M_k)P(D|\theta_k, M_k) = P(D, \theta_k|M_k) \qquad (۱۴)$$

سندیت از رابطه $Z = P(D|M_k) = \int P(D, \theta_k|M_k)d\theta_k$ به دست می‌آید. با تعریف

$$\beta_k = \sum_{s^{\leftarrow k}, s} [n(s^{\leftarrow k}s) + \alpha(s^{\leftarrow k}s)]$$

و توزیع Q به صورت

$$Q = \{q(s^{\leftarrow k}) = \frac{n(s^{\leftarrow k}) + \alpha(s^{\leftarrow k})}{\beta_k}, \quad q(s|s^{\leftarrow k}) = \frac{n(s^{\leftarrow k}s) + \alpha(s^{\leftarrow k}s)}{n(s^{\leftarrow k}) + \alpha(s^{\leftarrow k})}\}$$

می‌توان امید و واریانس $E(Q, P) = D[Q||P] + h_\mu[Q]$ را با مشتق گیری از لگاریتم تابع Z نسبت به β_k به دست آورد. که در آن P توزیع پارامترهای واقعی است: $P = \{p(s^{\leftarrow k}), p(s|s^{\leftarrow k})\}$. فرم مجانبی امید ریاضی $E(Q, P)$ به صورت زیر خواهد بود.

$$E_{post}[E(Q, P)] = H[q(s^{\leftarrow k})q(s|s^{\leftarrow k})] \qquad (۱۵)$$
$$-H[q(s^{\leftarrow k})] + \frac{۱}{۲\beta_k} | A |^k (| A | - ۱) + O(\frac{۱}{\beta_k^۲})$$

آنتروپی‌های مورد استفاده در معادله (۱۵) به صورت زیر تعریف می‌شوند.

$$H[q(s^{\leftarrow k})] = -\sum_{s^{\leftarrow k}} q(s^{\leftarrow k}) \log_۲ q(s^{\leftarrow k}) \qquad (۱٦)$$

$$H[q(s^{\leftarrow k})q(s|s^{\leftarrow k})] = \qquad (۱۷)$$
$$-\sum_{s^{\leftarrow k}, s} q(s^{\leftarrow k})q(s|s^{\leftarrow k}) \times \log_۲ q(s^{\leftarrow k})q(s|s^{\leftarrow k})$$

با توجه به معادله (۱۵) دو جمله اول نرخ آنتروپی را می‌سازند $h_\mu[Q] = H[q(s^{\leftarrow k})q(s|s^{\leftarrow k})] - H[q(s^{\leftarrow k})]$ و جمله آخر با آنتروپی نسبی شرطی بین توزیع میانگین پسین Q و توزیع واقعی P یکی است. از این رو می‌توان برای برآورد نرخ آنتروپی همگرایی $E_{post}[E(Q, P)]$ به نرخ آنتروپی $h_\mu[Q]$ را مورد بررسی قرار داد (برای جزییات بیشتر به [۳] مراجعه کنید.)

٦ مثال‌ها

در این قسمت نظریه‌های بیان شده در قسمت‌های قبل را در قالب یک مثال به صورت کاربردی بیان می‌کنیم. منبع داده‌های این مثال به فرآیند میانگین طلایی معروف است. این فرآیند یک زنجیر مارکف مرتبه اول است [۳]. فرآیند میانگین طلایی می‌تواند به صورت الگویی با دو وضعیت داخلی —آنها را A و B بخوانید— ثبت شود. هدف این قسمت، بررسی سه گام اصلی در استنباط برای تجزیه و تحلیل است. در ابتدا استنباط از یک

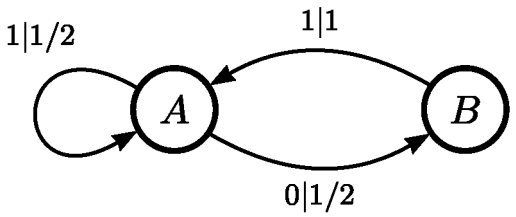
زنجیر مارکف مرتبه اول را برای توصیف برآورد پارامترهای الگو با عدم قطعیت، بررسی می‌کنیم. سپس، مقایسهٔ الگو را برای مرتبه‌های مختلف k بررسی می‌کنیم. این کار سبب می‌شود ساختار منبع داده‌ها را کشف کنیم. در نهایت، نرخ آنتروپی این منبع داده را برآورد می‌کنیم. برای این کار، از فرم شناخته شده منابع سود می‌بریم که هریک به وسیله یک ماتریس انتقال T که انتقال بین وضعیت‌های A و B را نشان می‌دهد، شرح داده می‌شود.

$$T = \begin{pmatrix} P(A \mid A) & P(B \mid A) \\ P(A \mid B) & P(B \mid B) \end{pmatrix}$$

وضعیت ویژه $\vec{\pi} = (P(A), P(B))$ توزیع مجانبی روی وضعیت‌های داخلی را شرح می‌دهد. ماتریس T می‌تواند به ماتریس‌های برچسب دار $T^{(s)}$ که شامل عناصری از T هستند که خروجی آنها نماد s است تقسیم شود. برای منابع داده‌های دودویی داریم: $T = T^{(0)} + T^{(1)}$. با استفاده از این ماتریس‌ها، متوسط احتمال کلمه‌ها را برای هر فرآیند دلخواه برآورد می‌کنیم. برای مثال، احتمال کلمهٔ ۰۱ با استفاده از رابطه $p(01) = \vec{\pi} T^{(0)} T^{(1)} \vec{\eta}$ به دست می‌آید. که در آن $\vec{\eta}$ یک بردار ستونی است که همه عناصرش ۱ هستند. برای داده‌های با اندازهٔ N، متوسط تعداد دفعاتی که یک کلمه با طول k+1 مشاهده می‌شود را به صورت مقابل برآورد می‌کنیم: $n(s^{\leftarrow k}s) = (N-k)p(s^{\leftarrow k}s)$. در برآورد نرخ آنتروپی برای مثال‌ها، از فرمول

$$h_\mu = -\sum_{v \in \{A,B\}} p(v) \sum_{s \in A} p(s \mid v) \log_2 p(s \mid v)$$

استفاده می‌کنیم. در این فرمول، $p(s \mid v) = T_v^{(s)}$ احتمال حرف s به شرط وضعیت داده شده v است و $p(v)$ احتمال مجانبی وضعیت v است که می‌تواند به صورتی که در بالا بحث شد، محاسبه شود.



شکل ۱. زنجیر مارکف برای فرآیند میانگین طلایی. یال‌ها با نماد خروجی و احتمال انتقال برچسب زده شده‌اند.

فرآیند میانگین طلایی می‌تواند به وسیلهٔ یک زنجیر مارکف مرتبه اول روی یک الفبای دودویی نمایش داده شود که در آن هرگز دو صفر متوالی دیده نمی‌شود. ماتریس‌های انتقال برچسب دار تعریف شده برای این منبع داده به صورت زیر هستند.

$$T^{(0)} = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix}, \qquad T^{(1)} = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix}$$

شکل ۱ یک نمایش گرافیکی از زنجیر مارکف پنهان منطبق با این فرآیند را نمایش می‌دهد. این شکل وجود یک رابطهٔ ساده بین وضعیت‌های A و B و نمادهای خروجی ۰ و ۱ را نشان می‌دهد. مشاهدهٔ ۰ نشان دهنده یک انتقال به وضعیت B

است و مشاهده ۱ دلالت بر یک انتقال به وضعیت A دارد و لذا این فرآیند یک زنجیر مارکف روی ۰ ها و ۱ ها ایجاد می‌کند. $p(A)$ برابر است با: $p(A) = p(۱) = \vec{\pi}T^{(۱)}\vec{\eta} = \frac{۲}{۳}$ و $p(B)$ برابر است با: $p(B) = p(۰) = \vec{\pi}T^{(۰)}\eta = \frac{۱}{۳}$ و لذا برای این فرآیند داریم $\vec{\pi} = [p(A), p(B)] = (\frac{۲}{۳}, \frac{۱}{۳})$ است. با این بردار و ماتریس‌های انتقال برچسب دار تعریف شده در بالا، متوسط تعداد دفعات رخداد هر کلمهٔ خواسته شده می‌تواند مطابق آنچه در بالا شرح داده شد به دست آید.

۱.۶ برآورد پارامترها

به منظور استنباط پارامترهای زنجیر مارکف برای فرآیند میانگین طلایی از متوسط تعداد دفعات تکرار هر کلمه در اندازه‌های متنوع N یعنی $n(s^{\leftarrow k}s)$ استفاده می‌کنیم. برای این کار روش دوم، یعنی بررسی توزیع حاشیه‌ای برای پارامترهای الگو را به کار می‌گیریم. با توجه به معادله (۹) چگالی پسین حاشیه‌ای برای توزیع دیریکله توزیع بتا است و برای پارامترهای $p(۰|۱)$ و $p(۱|۰)$ به ترتیب به صورت زیر است.

$$P(p(۰|۱)) = \frac{\Gamma[n(۱)+\alpha(۱)]}{\Gamma[n(۱۰)+\alpha(۱۰)]} \times \frac{p(۰|۱)^{n(۱۰)+\alpha(۱۰)-۱}}{\Gamma[n(۱)+\alpha(۱)-n(۱۰)-\alpha(۱۰)]} \times (۱-p(۰|۱))^{n(۱)+\alpha(۱)-n(۱۰)-\alpha(۱۰)-۱}$$

$$P(p(۱|۰)) = \frac{\Gamma[n(۰)+\alpha(۰)]}{\Gamma[n(۰۱)+\alpha(۰۱)]} \times \frac{p(۱|۰)^{n(۰۱)+\alpha(۰۱)-۱}}{\Gamma[n(۰)+\alpha(۰)-n(۰۱)-\alpha(۰۱)]} \times (۱-p(۱|۰))^{n(۰)+\alpha(۰)-n(۰۱)-\alpha(۰۱)-۱}$$

برای نمونه‌های با اندازه ۴۰۰ و ۲۰۰ و ۱۰۰ و N=۵۰، چگالی پسین حاشیه‌ای را برای پارامترهای $p(۰|۱)$ و $p(۱|۰)$ بدست می‌آوریم. برای این کار ابتدا مقادیر $n(s^{\leftarrow k}s)$ را از رابطه $n(s^{\leftarrow k}s) = (N-K)p(s^{\leftarrow k}s)$ به ازای $k = ۱$ می‌یابیم. برای مثال به ازای N=۵۰ مقادیر فوق به صورت جدول (۱) است:

جدول ۱. مقادیر $p(s^{\leftarrow k}s)$ و $n(s^{\leftarrow k}s)$ به ازای N=۵۰.

$n(۰۰) = ۰$	$p(۰۰) = \vec{\pi}T^{(۰)}T^{(۰)}\vec{\eta} = ۰$
$n(۰۱) = \frac{۴۹}{۳} = ۱۶/۳۳$	$p(۰۱) = \vec{\pi}T^{(۰)}T^{(۱)}\vec{\eta} = \frac{۱}{۳}$
$n(۱۰) = \frac{۴۹}{۳} = ۱۶/۳۳$	$p(۱۰) = \vec{\pi}T^{(۱)}T^{(۰)}\vec{\eta} = \frac{۱}{۳}$
$n(۱۱) = \frac{۴۹}{۳} = ۱۶/۳۳$	$p(۱۱) = \vec{\pi}T^{(۱)}T^{(۱)}\vec{\eta} = \frac{۱}{۳}$

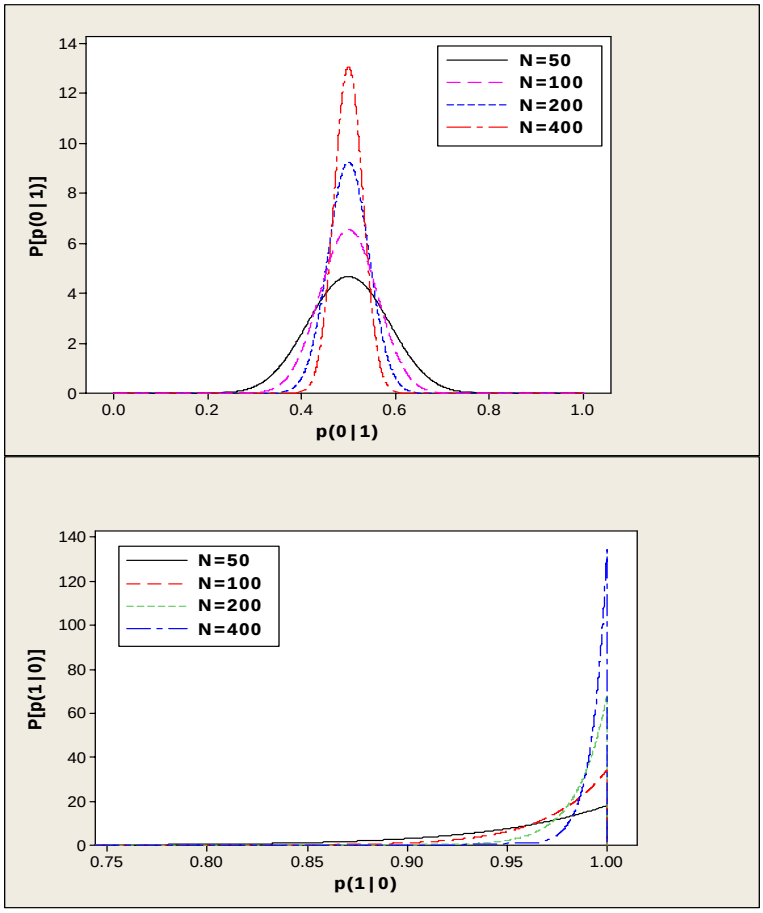
چگالی پسین حاشیه‌ای برای پارامتر $p(۰|۱)$ توزیع بتا با پارامترهای (۱۷/۳و۱۷/۳) است:

$$P(p(۰|۱)) = \frac{\Gamma(۳۴/۶۶)}{\Gamma(۱۷/۳۳)\Gamma(۱۷/۳۳)} \times p(۰|۱)^{۱۶/۳۳}(۱-p(۰|۱))^{۱۶/۳۳}$$

چگالی پسین حاشیه‌ای برای پارامتر $p(۱|۰)$ توزیع بتا با پارامترهای (۱و۱۷/۳) است:

$$P(p(۱|۰)) = \frac{\Gamma(۱۸/۳۳)}{\Gamma(۱۷/۳۳)\Gamma(۱)} \times p(۱|۰)^{۱۶/۳۳}(۱-p(۱|۰))^{۰}$$

چگالی پسین حاشیه‌ای برای پارامترهای $p(۰|۱)$ و $p(۱|۰)$ در شکل ۲ رسم شده است.



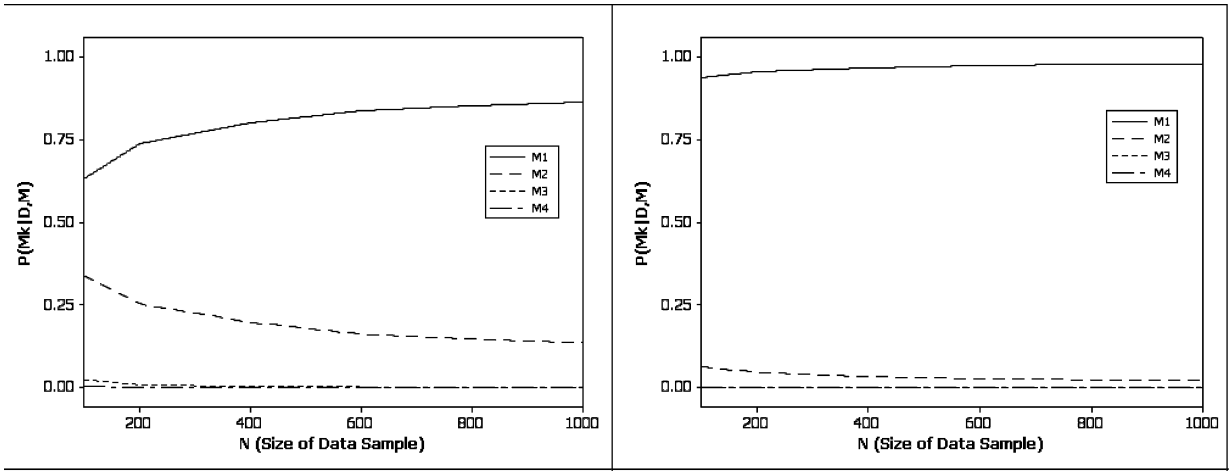
شکل ۲. نمودار استنباط پارامترهای الگوی M_1 برای فرآیند میانگین طلایی. برای هر نمونه داده به حجم N، چگالی پسین حاشیه ای برای پارامترهای مورد علاقه رسم شده است: در نمودار بالا $p(0|1)$ و در نمودار پایین $p(1|0)$ رسم شده است. مقادیر واقعی پارامترها $p(0|1) = \frac{1}{2}, p(1|0) = 1$ هستند.

نمودارها توزیع‌های ممکن برای پارامترهای الگو را برای هر N توصیف می‌کنند. همان‌طور که می‌بینیم با افزایش حجم داده‌ها به مقادیر صحیح $p(0 \mid 1) = \frac{1}{2}$ و $p(1 \mid 0) = 1$ همگرا می شوند. برآوردهای نقطه‌ای با واریانس‌های متناظر می‌تواند برای هر یک از پارامترها به دست آید، ولی این مقادیر خود می‌توانند گمراه کننده باشند. به هر حال، برآورد به دست آمده از به کارگیری میانگین و واریانس توزیع پسین توصیف واقعی‌تری از فرآیند استنباط نسبت به برآورد حداکثر درستنمایی، با خطای برآورد شده که تنها به وسیلهٔ تقریب گوسی از درستنمایی به دست می آید، ارائه می‌دهد. همان‌طور که شکل ۲ نشان می‌دهد، در حقیقت، تقریب گوسی از عدم قطعیت یک توصیف ناکارآمد از دانسته‌های ما زمانی است که پارامترهای زنجیر مارکف نزدیک حدود بالایی یا پائینی خود در ۰ و ۱ هستند. شاید نتیجه بخش ترین مجموعه از مقادیری که می‌توان به دست

آورد شامل میانگین توزیع پسین و ناحیهٔ اطمینان باشد، اینها با بیشترین دقت، عدم تقارن در عدم قطعیت پارامترهای الگو را توصیف خواهند کرد.

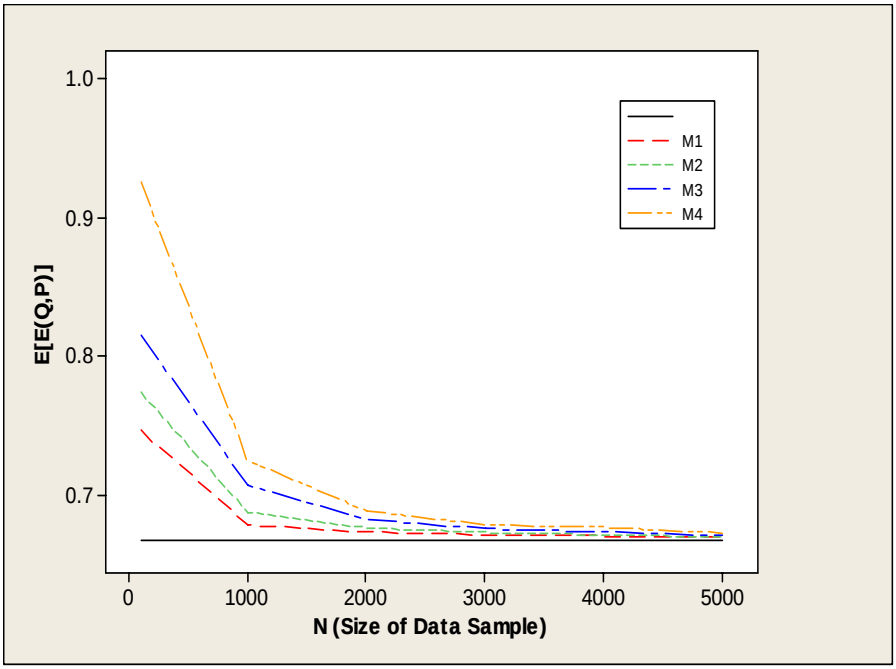
۲.۶ انتخاب رتبه الگو

همانطور که بیان کردیم فرآیند میانگین طلایی یک زنجیر مارکف مرتبه اول است ($k = ۱$). در نتیجه انتظار داریم مقایسهٔ الگو، این مرتبه را از میان احتمال‌های مورد بررسی انتخاب نماید. برای شرح این مسئله، مرتبه‌های $k = ۱, ..., ۴$ را بررسی می‌کنیم و مقایسهٔ الگو را با توزیع پیشین یکنواخت روی مرتبه‌ها (معادلهٔ ۱۱) و نیز با یک جریمه برای تعداد پارامترهای الگو (معادلهٔ ۱۲) انجام می‌دهیم. نتایج مقایسهٔ الگو در شکل ۳ نشان داده شده است.



شکل ۳. مقایسه الگو برای زنجیرهای مارکف از مرتبه $k = ۱, ..., ۴$ با استفاده از مقادیر متوسط فرآیند میانگین طلایی. حجم‌های نمونه از $N = ۱۰۰$ تا $N = ۱۰۰۰$ در گام‌هایی از $\Delta N = ۵$ برای تولید این نمودارها استفاده شده‌اند.

نمودار سمت چپ احتمال هر مرتبهٔ k را به صورت تابعی از حجم نمونه با به کارگیری توزیع پیشین یکنواخت، نشان می‌دهد. با این توزیع پیشین روی مرتبه‌ها، $M_۱$ با هر اندازه مناسب از داده‌ها انتخاب شده است. هرچند به نظر می‌رسد امکان بیش برازش برای حجم کم داده‌ها ($N \leq ۱۰۰$) وجود دارد. نمودار سمت راست احتمال الگو را با توزیع پیشین با جریمه روی مرتبه k نشان می‌دهد. این توزیع پیشین بیش برازش شده را در حجم‌های کوچک داده‌ها برطرف می‌کند. البته این مطلب به معنی بهتر بودن این توزیع پیشین نیست، بلکه در حقیقت، مقایسهٔ الگوی بیزی با توزیع پیشین یکنواخت یک کار مؤثر با به کارگیری حجم نمونه نسبتاً کوچک است.



شکل ۴. همگرایی $E_{post}[E[Q,P]]$ به نرخ واقعی آنتروپی $h_\mu = \frac{2}{3}$ افقی نشان داده شده است) (در شکل با خ برای فرآیند میانگین طلایی. همانطور که در معادله (۱۵) شرح داده شد، آنتروپی نسبی شرطی $D[Q||P] \to 0$ و این همگرایی در همگرایی $h_\mu[Q]$ به نرخ واقعی آنتروپی نتیجه می‌شود.

۳.۶ برآورد نرخ آنتروپی

حال می‌توانیم همگرایی میانگین $E(Q,P) = D[Q||P] + h_\mu[Q]$ را به نرخ درست آنتروپی برای فرآیند میانگین طلایی شرح دهیم. این همگرایی را برای همهٔ مرتبه‌های $k = 1,\ldots,4$ که در قسمت قبل بحث شد، نشان می‌دهیم. این مثال نشان می‌دهد که همهٔ مرتبه‌های بزرگتر یا مساوی $k = 1$ به نرخ آنتروپی میل می‌کنند. هرچند همگرایی به مقادیر واقعی برای مرتبه‌های بالاتر به خاطر بزرگ بودن مقدار اولیهٔ $D[Q||P]$ دربرگیرنده داده‌های بیشتری است. این مقدار بزرگتر به دلیل تعداد بزرگتر پارامترها برای زنجیرهای مارکف مراتب بالاتر است. در ارزیابی مقدار $D[Q||P] + h_\mu[Q]$ برای نمونه‌های با اندازه متفاوت، انتظار داریم که برآورد میانگین توزیع پسین Q به توزیع واقعی P همگرا شود. در نتیجه باید آنتروپی نسبی شرطی با افزایش N به صفر میل کند. برای فرآیند میانگین طلایی، مقدار واقعی نرخ آنتروپی برابر با $h_\mu = \frac{2}{3}$ است. شکل ۴ همگرایی مورد انتظار از میانگین $E(Q,P)$ به نرخ واقعی آنتروپی را نشان می‌دهد. نتیجه‌ای که از مقایسهٔ الگو در بخش قبل گرفتیم، می‌تواند در برآورد نرخ آنتروپی نیز مورد استفاده قرار گیرد. همان‌طور که در شکل ۳ دیدیم، دامنه‌هایی از اندازه نمونه N وجود دارند که در آن‌ها احتمال مرتبه‌های

$k = ۱, ۲$ هر دو غیر صفر هستند. در اصل، برآورد h_μ باید با وزن‌دار کردن مقادیر به دست آمده برای هر k به وسیلهٔ احتمال مرتبه مطابق با آن $P(M_k \mid D, M)$ مشخص شود. همانطور که از شکل ۴ می‌بینیم، برآوردهای نرخ آنتروپی برای $k = ۱, ۲$ در این محدوده از N نیز بسیار مشابه هستند. در نتیجه، این روش اضافی نبایستی تأثیر بزرگی بر برآورد نرخ آنتروپی داشته باشد.

۴.٦ نتیجه گیری

در این مقاله با استفاده از مقاله [۳]، استنباط بیزی از الگوهای زنجیر مارکف مرتبهٔ k را که شامل برآورد پارامترهای الگو برای یک k داده شده به همراه مقایسهٔ الگو بین مرتبه‌ها و برآورد تصادفی بودن با استفاده از نرخ‌های آنتروپی بود، بررسی کردیم. در اغلب روش‌های استنباط، این سه وجه به صورت مجزا ولی در ارتباط با هم به کار می‌روند. در این مقاله نیز آنها را بسیار وابسته به هم یافتیم. برآورد پارامترهای الگو بدون در نظر گرفتن اینکه آیا الگوی صحیح انتخاب شده است یا نه، در بهترین حالت نیز گمراه کننده است. مقایسهٔ الگو به وسیله مقایسهٔ مرتبه‌های متنوع k در حدود کلاس الگو راهی به سوی حل این مسئله است. در نهایت، برآورد تصادفی بودن با استفاده از نرخ‌های آنتروپی، اطلاع بیشتری دربارهٔ رابطه بین ساختار و تصادفی بودن تولید می‌کند. این بینش‌ها علی رغم داده‌های خارج از کلاس، توان ترکیب این سه روش را به صورت یک ابزار مؤثر برای بررسی ساختار و تصادفی بودن در رشته‌های متناهی از داده‌های گسسته، شرح می‌دهند.

مراجع

[1] Thornburg, H. (2000), *Introduction to Bayesian Statistics*, Standford, California: Standford University.

[2] Anderson, T.W. and Goodman, L.A. (1957), Statistical inference about Markov chains, *The Annals Mathematicals Statistics*, 28, 89-110.

[3] Strelioff, C.C., Crutchfield, J.P. and Hubler, A.W. (2007), Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling, *Phys. Rev. E.*, **76**, 011106.1-011106.14.

[4] Chatfield, C. (1973), Statistical inference regarding markov chain models, *Applied Statistics*, 22, 7-20.

[5] Liu, J.S. and Lawrece, C.E. (1999), Bayesian inference on biopolymer models, *Bioinformatics*, 15, 38-52.

[6] Cover, T.M. and Thomas, J.A. (1991), *Elements of Information Theory*, Wiley, New York.

[7] Shannon, C.E. (1948), A mathematical theory of communication, *Bell System Technical Journal*, 27, 379-423.

[8] MacKay, D.J.C. and Peto, L.C.B. (1994), A hierarchical dirichlet language model, *Nat. Lang. Eng.*, 1, 1-19.