مجله علوم آماری، پاییز و زمستان ۱۳۸۷
جلد ۲، شماره ۲، ص ۲۲۹-۲۴۳

برآورد بیزی تابع تاوان در آزمون همگنی مدل‌های آمیخته

رحمان فرنوش[۱]، افشین فلاح[۲]، آرزو حاج رجبی[۱]
[۱]گروه آمار، دانشگاه علم و صنعت
[۲]گروه آمار، دانشگاه بین‌المللی امام خمینی

تاریخ دریافت: ۱۳۸۷/۷/۱۷ تاریخ آخرین بازنگری: ۱۳۸۷/۱۲/۲۰

چکیده : برای آزمون فرضیه همگنی مدل‌های آمیخته، معمولاً از آزمون نسبت درستنمایی اصلاح شده که مبتنی بر افزودن یک تابع تاوان مناسب به تابع لگ درستنمایی می‌باشد، استفاده می‌شود. کارایی این آزمون به شدت تحت تأثیر شکل تابع تاوان انتخابی است. انتخاب تابع تاوان در این نوع آزمون معمولاً بر اساس پرهیز از پیچیدگی و میسر بودن برآورد پارامترها صورت می‌پذیرد، که لزوماً نتایج مطلوبی بدنبال ندارد. در این مقاله یک تابع تاوان جامع در نظر گرفته شده است، که دارای یک پارامتر تعیین کنندهٔ شکل است. سپس پارامتر تعیین کنندهٔ شکل این تابع تاوان و پارامترهای مدل آمیخته با در نظر گرفتن توزیع‌های پیشین مناسب برای آن‌ها با استفاده از رهیافت بیزی، بصورت پسینی برآورد شده‌اند. نشان داده شده است که رهیافت بیزی پیشنهادی در براورد پارامترهای مدل، در مقایسه با رهیافت بسامدی، به مراتب کارایی مطلوب‌تری دارد. این کارایی خصوصاً در شرایط شناخت ناپذیری توزیع آمیخته که روش‌های بسامدی کارایی اندکی دارند، بیشتر است.

واژه‌های کلیدی : آزمون نسبت درستنمایی، تابع تاوان، الگوریتم EM، زنجیرهای مارکف مونت کارلوئی.

آدرس الکترونیک مسئول مقاله: رحمان فرنوش، rfarnoosh@iust.ac.ir
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲F۱۵

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

۱ مقدمه

توزیع‌های آمیخته به دلیل انعطاف پذیر بودن در توصیف بسیاری از پدیده‌های تصادفی مفید هستند. پس از مقاله کارل پیرسون در دهۀ هشتاد این توزیع‌ها بطور گسترده‌ای در تحقیقات کاربردی و در زمینه‌های متفاوتی مانند ژنتیک، زیست شناسی، اقتصاد و غیره بکار گرفته می‌شوند. کاربرد توزیع‌های آمیخته معمولاً در مواردی است که جامعه‌ی آماری ناهمگن و ترکیبی از چند زیر جامعه است. در چنین حالتی مشاهداتی که از این جامعه بدست می‌آیند، می‌توانند با احتمال معینی به هر یک از این زیر جامعه‌ها تعلق داشته باشند. روش‌های مختلفی برای آزمون تعداد مؤلفه‌های یک توزیع آمیخته وجود دارد، که از آن جمله می‌توان به روش‌هایی بر پایه نظریه اطلاع، فاصله تاوانیده، روش لگ درستنمایی تاوانیده و روش بیزی اشاره نمود. آکائیک (۱۹۷۳) به کمک نظریه اطلاع، معیار اطلاع آکائیک را برای انتخاب تعداد مؤلفه‌های مدل آمیخته، از طریق مینیمم ساختن فاصله کولبک–لایبلر بین توزیع جامعه و توزیع مدل‌های پیشنهادی مورد مطالعه قرار داد. چن و کال فلیسک (۱۹۹۶) روشی را بر پایه مینیمم ساختن فاصلۀ تاوانیده بین تابع توزیع تجربی و تابع توزیع تجمعی برازش داده شده، ارائه دادند. چن و خلیلی (۲۰۰۶) روشی را بر پایه لگ درستنمایی تاوانیده، برای انتخاب تعداد مؤلفه‌های مدل آمیخته معرفی کردند. ریچاردسون و گرین (۱۹۹۷) انتخاب تعداد مؤلفه‌های مدل آمیخته را از دیدگاه بیزی مورد مطالعه قرار دادند. مارین و همکاران (۲۰۰۵) دشواری‌های تحلیل بیزی مدل‌های آمیخته را بررسی نموده و نشان دادند که وقتی توزیع جامعه آمیخته است، کاربست روشهای معمول زنجیر مارکف مونت کارلویی برای نمونه‌گیری از توزیع پسین پارامترها با دشواری‌های زیادی روبرو است. آزمون فرضیۀ همگن بودن جامعه توسط محققان بسیاری مورد توجه قرار گرفته است. در حالت خاص، برای انتخاب یک مدل همگن در مقابل یک مدل آمیخته با دو مؤلفه، هارتیگان (۱۹۸۵) استفاده از آزمون نسبت درستنمایی را پیشنهاد نمود. ولی استفاده از این روش به دلیل پیچیدگی توزیع حدی آمارۀ آزمون نسبت درستنمایی، عملاً امکان پذیر نیست. چن (۱۹۹۸) منابع بی‌نظمی مؤثر بر توزیع حدی آمارۀ نسبت درستنمایی را مورد مطالعه قرار داد و بر این اساس آزمون نسبت درستنمایی اصلاح شده را معرفی نمود، که در آن با افزودن یک تابع تاوان به تابع لگ درستنمایی، توزیع حدی ساده‌ای برای آماره‌ی آزمون نسبت درستنمایی بدست می‌آید. یکی از محدودیّت‌های این روش آن است که استفاده از آزمون نسبت درستنمایی اصلاح شده، منوط به برقراری برخی شرایط نظم است. بعلاوه توان این آزمون به شدت تحت تاثیر تابع تاوان انتخابی است. از اینرو لی و همکاران (۲۰۰۸) آزمون EM را بر اساس شکل دیگری از

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

تابع تاوان پیشنهادی توسط چن و همکاران (۲۰۰۱)، که مستقل از فرضیات لازم برای آزمون نسبت درستنمایی اصلاح شده است، مطرح نمودند.

در این مقاله آزمون فرض همگنی جامعه با در نظر گرفتن فرم جامعی برای تابع تاوان که به یک پارامتر شکل وابسته است، از دیدگاه بیزی مورد مطالعه قرار گرفته است. برای این منظور با در نظر گرفتن پیشین‌های مناسب، پارامتر تعیین کنندهٔ شکل تابع تاوان و پارامترهای مدل آمیخته بصورت پسینی برآورد شده‌اند. در بخش ۲ آزمون نسبت درستنمایی و شکل اصلاح شدهٔ آن معرفی و مشکلات این نوع آزمونها مورد بررسی قرار گرفته است. در بخش ۳ چگونگی برآورد پارامترهای مدل با استفاده از الگوریتم EM شرح داده شده است. مسئلهٔ انتخاب تابع تاوان در بخش ۴ مورد بحث قرار گرفته و روشی برای برآورد پارامتر تعیین کنندهٔ شکل تابع تاوان به کمک رهیافت بیزی، پیشنهاد شده است. در بخش ۵، کارایی آزمون نسبت درستنمایی و روش بیزی پیشنهادی با استفاده از یک مطالعهٔ شبیه‌سازی مورد ارزیابی قرار گرفته است.

۲ آزمون نسبت درستنمایی

متغیر تصادفی Y دارای توزیع آمیخته است، هرگاه تابع چگالی یا جرم احتمال آن بصورت

$$p_Y(y;\theta,p)=\sum_{j=1}^{k}p_jf_j(y|\theta_j),\ y\in\mathcal{Y},\ 0\le p_j\le 1,\ j=1,\ldots,k,\ \sum_{j=1}^{k}p_j=1,$$

باشد، که درآن $\mathcal{Y}$ تکیه‌گاه متغیر تصادفی Y را نشان می‌دهد و $\theta=(\theta_1,\ldots,\theta_k)$، که θ_j پارامتر مؤلفه‌ی jام و $p=(p_1,\ldots,p_k)$، که p_j نسبت آمیخته مؤلفه‌ی j ام نامیده می‌شود. هنگامی که از یک توزیع آمیخته نمونه‌ای مشاهده می‌شود، از آنجا که مشخص نیست هر مشاهده مربوط به کدام زیر جامعه می‌باشد، مشاهدات بدست آمده را داده‌های ناقص گویند. در اینصورت تابع لگاریتم درستنمایی داده‌های ناقص بصورت

$$\ell_n(\theta,p)=\sum_{i=1}^{n}\log(\sum_{j=1}^{k}p_jf_j(y_i|\theta_j)),$$

است. داده‌های ناقص را می‌توان با در نظر گرفتن مجموعه‌ای از متغیرهای نشانگر که تعلق مشاهدات به زیر مجموعه‌ها را مشخص می‌سازند، به داده‌های کامل تبدیل نمود. تابع لگاریتم

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

درستنمایی داده‌های کامل که معمولاً برای انجام استنباط مناسبتر می‌باشد، بصورت

$$\ell_n^c(z,\theta,p)=\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\{\log p_j+\log f_j(y_i|\theta_j)\},$$

است. درستنمایی داده‌های کامل، تابعی از متغیرهای پنهان $Z=(z_1,\ldots,z_n)$ و مشاهدات می‌باشد، که در آن $z_i=(z_{i1},\ldots,z_{ik})$ و $z_{ij}=1$ نشان دهنده تعلق مشاهده iام به مؤلفه‌ی jام است. فرض کنید $Y_1,\ldots,Y_n$ نمونه‌ای تصادفی از توزیع آمیخته

$$pf_Y(y|\theta_1)+(1-p)f_Y(y|\theta_2), \tag{1}$$

باشد، که در آن $\theta_1,\theta_2\in\Theta$ و Θ زیر مجموعه‌ای فشرده از خط حقیقی است. هدف آزمون

$$\begin{cases}\mathcal{H}_0: p(1-p)(\theta_1-\theta_2)=0\\ \mathcal{H}_1: p(1-p)(\theta_1-\theta_2)\neq 0,\end{cases} \tag{2}$$

است، که در آن فرض صفر به معنی همگنی جامعه و فرض مقابل به معنی تشکیل جامعه از دو زیر جامعه ناهمگن مطابق رابطه (۱) است. تابع لگاریتم درستنمایی را می‌توان بصورت

$$\ell_n(p,\theta_1,\theta_2)=\sum_{i=1}^{n}\log\{pf(y_i|\theta_1)+(1-p)f(y_i|\theta_2)\},$$

نوشت. اگر $\hat{\theta}_0$ و $(\hat{p},\hat{\theta}_1,\hat{\theta}_2)$ به ترتیب ماکسیمم کننده تابع درستنمایی تحت فرض صفر و مقابل باشند، در اینصورت آماره آزمون نسبت درستنمایی بصورت

$$R_n=2\{\ell_n(\hat{p},\hat{\theta}_1,\hat{\theta}_2)-\ell_n(0/5,\hat{\theta}_0,\hat{\theta}_0)\},$$

می‌باشد و مقادیر بزرگ این آماره منجر به رد فرض صفر می‌شوند. چن (۱۹۹۸) نشان داد دو منبع بی‌نظمی که توزیع حدی آماره آزمون نسبت درستنمایی را به طور قابل ملاحظه‌ای پیچیده می‌سازند، یکی قرار داشتن نقاط $p=0$ یا $p=1$ در محدودهٔ فرض صفر و دیگری شناخت ناپذیر بودن توزیع آمیخته (۱) تحت فرض صفر است. وی پیشنهاد نمود برای برطرف کردن این مشکل، تابع تاوانی برحسب p به صورت $T(p)$ که در شرایط

$$\lim_{p\to 0\, or\, 1} T(p)=-\infty \quad , \quad \arg\max_{p\in[0,1]} T(p)=0/5, \tag{3}$$

صدق کند، به آماره آزمون نسبت درستنمایی افزوده شود. در اینصورت تابع لگاریتم درستنمایی تاوانیده بصورت $T\ell_n(p,\theta_1,\theta_2)=\ell_n(p,\theta_1,\theta_2)+T(p)$ است و آماره آزمون نسبت درستنمایی بصورت $M_n=2\{T\ell_n(\hat{p},\hat{\theta}_1,\hat{\theta}_2)-T\ell_n(0/5,\hat{\theta}_0,\hat{\theta}_0)\}$ اصلاح می‌شود.

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

استفاده از این تابع تاوان موجب می‌شود مقادیر برازش داده شده p تحت درستنمایی اصلاح شده دور از صفر و یک واقع شوند. چن (۱۹۹۸) نشان داد که اگر برخی شرایط نظم روی هسته چگالی برقرار باشند، آنگاه توزیع مجانبی آماره آزمون نسبت درستنمایی اصلاح شده، آمیخته‌ای از دو توزیع با وزن‌های یکسان بصورت $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ می‌باشد، که در آن χ^2_0 توزیع تباهیده در نقطه صفر و χ^2_1 توزیع کای دو با درجه آزادی ۱ را نشان می‌دهند. نکته‌ی قابل توجه آن است که این توزیع حدی به فرم تابع تاوان انتخابی وابسته نیست.

۳ برآورد پارامترها با الگوریتم EM

برای یافتن برآورد ماکسیمم درستنمایی پارامترها در توزیع‌های آمیخته، معمولاً از الگوریتم تکراری EM استفاده می‌شود. که شامل دو مرحله امیدگیری E و ماکسیمم سازی M است. فرض کنید $Y_1, \ldots, Y_n$ نمونه‌ای تصادفی از توزیع آمیخته (۱) باشد، در این صورت تابع لگاریتم درستنمایی داده‌های کامل بصورت

$$\begin{aligned}\ell_n^c((\theta_1,\theta_2),p) &= \sum_{i=1}^{n}[z_{i1}\{\log(p)+\log(f(y_i|\theta_1)\} \\ &\quad +(1-z_{i1})\{\log(1-p)+\log(f(y_i|\theta_2)\}], \end{aligned} \qquad (۴)$$

است. در مرحله E، امید ریاضی

$$\begin{aligned} z_{i1}^{(t)} &= E(z_{i1}|y_i,\theta^{(t-1)},p^{(t-1)}) \\ &= \frac{p^{(t-1)}f(y_i|\theta_1^{(t-1)})}{p^{(t-1)}f(y_i|\theta_1^{(t-1)})+(1-p^{(t-1)})f(y_i|\theta_2^{(t-1)})}, \quad i=1,\ldots,n, \end{aligned} \qquad (۵)$$

محاسبه و جایگزین $z_{ij}, i = 1, \ldots, n$ می‌شود. در مرحله M، با قرار دادن مقادیر (۵) در تابع لگ درستنمایی داده‌های کامل و ماکسیمم کردن این تابع نسبت به پارامترهای مدل، داریم $p^{(t)} = \frac{1}{n}\sum_{i=1}^{n} z_{i1}^{(t)}$ و

$$\begin{aligned}\theta_1^{(t)} &= \arg\max_{\theta_1\in\Theta}\sum_{i=1}^{n}\{z_{i1}^{(t)}\log(f(y_i|\theta_1)\}, \\ \theta_2^{(t)} &= \arg\max_{\theta_2\in\Theta}\sum_{i=1}^{n}\{(1-z_{i1}^{(t)})\log(f(y_i|\theta_2)\}.\end{aligned}$$

در این صورت، تکرار مراحل E و M تا حصول همگرایی، برآوردهای ماکسیمم درستنمایی پارامترها را بدست می‌دهد. برآورد p به انتخاب تابع تاوان بستگی دارد، چن و همکاران

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

(۲۰۰۱) تابع تاوانی به صورت

$$T(p) = C \log(۴p(۱-p)), \qquad (۶)$$

را پیشنهاد نمودند، که در آن C ثابتی مثبت و تأثیر گذار بر میزان اصلاح تابع درستنمایی است. لی و همکاران (۲۰۰۸) نیز تابع تاوانی به فرم

$$T(p) = C^* \log(۱ - |۱ - ۲p|), \qquad (۷)$$

را مورد استفاده قرار دادند. توابع تاوان (۶) و (۷) در شرایط (۳) صدق می‌کنند و به سادگی می‌توان نشان داد که مقدار p در مرحله‌ی M از تکرار tام الگوریتم EM، برای این توابع به ترتیب بصورت $p^{(t+۱)} = \frac{\sum_{i=۱}^{n} z_{i۱}^{(t)} + C}{۲C+n}$ و

$$p^{(t)} = \begin{cases} \min\{\frac{\sum_{i=۱}^{n} z_{i۱}^{(t)} + C^*}{n+C^*}, ۰/۵\} & \frac{\sum_{i=۱}^{n} z_{i۱}^{(t)}}{n} < ۰/۵, \\ ۰/۵ & \frac{\sum_{i=۱}^{n} z_{i۱}^{(t)}}{n} = ۰/۵, \\ \max\{\frac{\sum_{i=۱}^{n} z_{i۱}^{(t)}}{n+C^*}, ۰/۵\} & \frac{\sum_{i=۱}^{n} z_{i۱}^{(t)}}{n} < ۰/۵, \end{cases}$$

بدست می‌آیند.

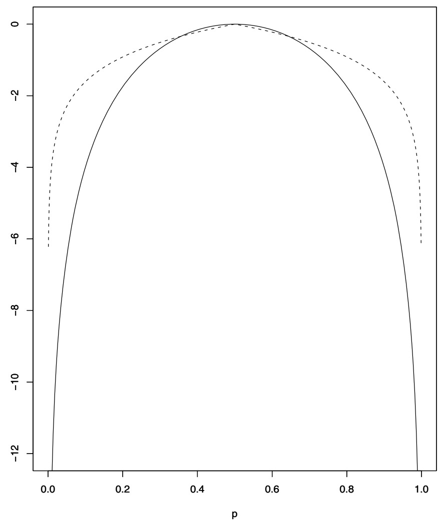
۴ انتخاب تابع تاوان

گرچه توزیع حدی آماره آزمون نسبت درستنمایی به فرم تابع تاوان انتخابی وابسته نیست، اما فرم تابع تاوان انتخاب شده بر توان آزمون تأثیر گزار است. آزمون اصلاح شده‌ای که از افزودن تابع تاوان (۶) به تابع لگاریتم درستنمایی حاصل می‌شود، مشکلات آزمون نسبت درستنمایی معمول را تاحدودی مرتفع می‌سازد. با این وجود گاهی حتی با وجود مشاهدات مشهود از توزیع آمیخته، آزمون نسبت درستنمایی اصلاح شده قادر به رد فرض $\mathcal{H}_۰$ نیست. دلیل این امر آن است که تابع تاوان (۶) به ازای نسبت‌های آمیخته نزدیک صفر یا یک، تاوان زیادی را روی تابع لگاریتم درستنمایی اعمال می‌کند. به همین دلیل به تابع تاوان معقول‌تری نیاز است، به نحوی که برای آن توان آزمون نسبت درستنمایی اصلاح شده، حتی در حالتی که نسبت‌های آمیخته به صفر و یک نزدیک هستند، افزایش یابد. تابع تاوان (۷) به این منظور پیشنهاد شده است. به سادگی ملاحظه می‌شود که بین دو تابع تاوان (۶) و (۷) نابرابری

$$\log(۱ - |۱ - ۲p|) \leq \log(۱ - |۱ - ۲p|^۲) = \log(۴p(۱-p)),$$

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

برقرار است، که به ازای مقادیر نزدیک به ۰/۵ نسبت آمیخته، این نابرابری به برابری تقریبی $\log(۱-|۱-۲p|)\approx -|۱-۲p|$ تبدیل می‌شود. بنابراین توابع تاوان (۶) و (۷) به ازای مقادیر p نزدیک به ۰/۵ تقریباً معادل‌اند، اما به ازای مقادیر p نزدیک به صفر یا یک، تفاوت قابل ملاحظه‌ای بین آنها وجود دارد. شکل ۱ نمودار این دو تابع تاوان را به ازای $C^*=۱$ نشان می دهد. ملاحظه می‌شود که تابع تاوان (۶) در نقطه $p=۰/۵$ تقریباً مسطح می‌باشد و تاوانی را اعمال نمی‌کند، در حالی که تابع تاوان (۷) علاوه بر مرتفع‌تر بودن، به ازای مقادیر نزدیک به صفر و یک نیز تاوان بسیار کمتری را لحاظ می‌نماید.



شکل ۱: نمودار توابع تاوان (۶) (خط) و (۷) (نقطه چین).

با وجود اینکه تابع تاوان پیشنهادی لی و همکاران (۲۰۰۸)، دارای مزیت‌هایی می‌باشد و کارایی آزمون نسبت درستنمایی را افزایش می‌دهد، اما دلیلی مبنی بر بهینه بودن هیچ یک از این توابع تاوان وجود ندارد. بسادگی ملاحظه می‌شود که توابع تاوان (۶) و (۷) حالات خاصی از تابع تاوان کلی

$$g(p,h)=C\log(۱-|۱-۲p|^h)\ ,\quad ۰<h\leq ۲ \tag{۸}$$

هستند، که به ازای مقادیر $h=۱$ و $h=۲$ حاصل می‌شوند. انتخاب مقدار h به شکل تابع

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

تاوان و در نتیجه به استنباط‌های حاصل از آن به شدت تأثیر گذار است. در رهیافت بسامدی انتخاب مقادیر دیگری برای h به پیچیدگی تابع تاوان (۸) نتیجه می‌شود، که آن نیز به نوبه خود منجر به دشواری ماکسیمم سازی در مرحله M از الگوریتم EM می‌شود. با توجه به مشکلات رهیافت بسامدی در تعیین پارامترهای مدل و تابع تاوان، در این بخش این پارامترها در چارچوب رهیافت بیزی برآورد می‌شوند. برای این منظور ابتدا برای پارامترهای مورد نظر، توزیع‌های پیشین مناسب در نظر گرفته و سپس برآورد این پارامترها بصورت پسینی محاسبه می‌شوند. در حالت کلی، دامنهٔ تغییرات پارامتر h در تابع تاوان (۸) مجموعه اعداد حقیقی مثبت است. با این وجود میزان تاوانی که این تابع اعمال می‌کند با بزرگ شدن h به صفر میل می‌نماید، به نحوی که برای مقادیر $h > ۱$ میزان تاوان اعمال شده توسط این تابع اندک بوده و برای مقادیر $h > ۲$ عملاً هیچ تاوانی اعمال نمی‌شود. از این رو برخی محققان مانند لی و همکاران (۲۰۰۸) مقداری در بازه‌ی $[۰, ۱]$ را برای این پارامتر پیشنهاد کرده‌اند. در این مقاله از توزیع $U(۰, ۱)$ به عنوان توزیع پیشین برای پارامتر h استفاده شده است. به منظور حفظ بی طرفی، توزیع پیشین برای نسبت آمیخته p، $U(۰, ۱)$ در نظر گرفته شده است. چون توابع هسته مختلف تشکیل دهندهٔ یک توزیع آمیخته، دارای پارامترهای متفاوتی هستند، برای انجام تحلیل بیزی لازم است برای هر یک از این پارامترها نیز توزیع پیشین مناسبی لحاظ شوند. اگر $Y_۱, ..., Y_n$ نمونه‌ای تصادفی از توزیع آمیخته متناهی دو مؤلفه‌ای نرمال باشد، در این صورت تابع درستنمایی تاوانیده که از ضرب تابع درستنمایی در تابع تاوان حاصل می‌شود، بصورت

$$L(y|\mu_۱, \mu_۲, p, h) \quad \propto \quad p^{n_۱}(۱-p)^{n_۲} e^{-\frac{n_۱}{۲}(\bar{y}_۱-\mu_۱)^۲} e^{-\frac{n_۲}{۲}(\bar{y}_۲-\mu_۲)^۲}(۱-|۱-۲p|^h)$$

است، که در آن $n_۱$ تعداد مشاهدات منتسب به مؤلفه اول، $n_۲$ تعداد مشاهدات منتسب به مؤلفه دوم، $\bar{y}_۱$ میانگین مشاهدات منتسب به مؤلفه اول و $\bar{y}_۲$ میانگین مشاهدات منتسب به مؤلفه دوم است. با فرض اینکه h و p بصورت پیشینی مستقل هستند و با در نظر گرفتن توزیع‌های پیشین نرمال $\mu_j \sim N(\delta, \frac{۱}{\lambda}), j = ۱, ۲$ با $\delta \in R, \lambda > ۰$، برای پارامترهای مکانی دو مؤلفهٔ توزیع آمیخته نرمال، می‌توان نوشت

$$\begin{aligned}
\pi(\mu_۱, \mu_۲) &\propto e^{-\frac{\lambda}{۲}\{(\mu_۱-\delta)^۲+(\mu_۲-\delta)^۲\}}, \\
\pi(h, p) &= \pi(h)\pi(p) = ۱, \\
\pi(\mu_۱, \mu_۲, h, p) &\propto e^{-\frac{\lambda}{۲}\{(\mu_۱-\delta)^۲+(\mu_۲-\delta)^۲\}}, \\
\pi(\mu_۱, \mu_۲, h, p|y) &= f(y|\mu_۱, \mu_۲, p, h)\pi(\mu_۱, \mu_۲, h, p).
\end{aligned}$$

از این رو توزیع پسین را می‌توان بصورت

$$\pi(\mu_1,\mu_2,h,p|y) \propto p^{n_1}(1-p)^{n_2}e^{-\frac{n_1}{2}(\bar{y}_1-\mu_1)^2}e^{-\frac{n_2}{2}(\bar{y}_2-\mu_2)^2}(1-|1-2p|^h) \\ e^{-\frac{\lambda}{2}\{(\mu_1-\delta)^2+(\mu_2-\delta)^2\}},$$

نوشت. به همین ترتیب، اگر $Y_1,...,Y_n$ یک نمونه تصادفی از توزیع آمیخته متناهی دو مؤلفه‌ای پواسن باشد، در این صورت تابع درستنمایی تاوانیده از ضرب تابع درستنمایی در تابع تاوان بصورت

$$L(y|\lambda_1,\lambda_2,p,h) \propto p^{n_1}(1-p)^{n_2}\exp\{-n_1\lambda_1-n_2\lambda_2+n_1\bar{y}_1\ln\lambda_1+n_2\bar{y}_2\ln\lambda_2\} \\ (1-|1-2p|^h),$$

حاصل می‌شود، که در آن n_1 تعداد مشاهدات منتسب به مؤلفه اول، n_2 تعداد مشاهدات منتسب به مؤلفه دوم، $\bar{y}_1$ میانگین مشاهدات منتسب به مؤلفه اول و $\bar{y}_2$ میانگین مشاهدات منتسب به مؤلفه دوم است. با فرض اینکه h و p بصورت پیشینی مستقل هستند و با در نظر گرفتن توزیع‌های پیشین گاما $\lambda_j \sim \Gamma(\alpha_j,\beta_j),\ \ j=1,2$، با $\alpha_j>0,\beta_j>0$، برای پارامترهای λ در دو مؤلفه‌ای توزیع آمیخته‌ی پواسن، می‌توان نوشت

$$\begin{aligned}
\pi(\lambda_1,\lambda_2) &\propto \lambda_1^{(\alpha_1-1)}e^{-\frac{1}{\beta_1}\lambda_1}\lambda_2^{(\alpha_2-1)}e^{-\frac{1}{\beta_2}\lambda_2},\\
\pi(h,p) &= \pi(h)\pi(p)=1,\\
\pi(\lambda_1,\lambda_2,h,p) &\propto \lambda_1^{(\alpha_1-1)}e^{-\frac{1}{\beta_1}\lambda_1}\lambda_2^{(\alpha_2-1)}e^{-\frac{1}{\beta_2}\lambda_2},\\
\pi(\lambda_1,\lambda_2,h,p|y) &= f(y|\lambda_1,\lambda_2,p,h)\pi(\lambda_1,\lambda_2,h,p),
\end{aligned}$$

از این رو توزیع پسین را می‌توان بصورت

$$\pi(\lambda_1,\lambda_2,h,p|y) \propto p^{n_1}(1-p)^{n_2}e^{\{-n_1\lambda_1-n_2\lambda_2+n_1\bar{y}_1\ln\lambda_1+n_2\bar{y}_2\ln\lambda_2\}} \\ (1-|1-2p|^h)\lambda_1^{(\alpha_1-1)}e^{-\frac{1}{\beta_1}\lambda_1}\lambda_2^{(\alpha_2-1)}e^{-\frac{1}{\beta_2}\lambda_2},$$

نوشت. تحلیل بیزی مدل‌های آمیخته به دلیل پیچیدگی ذاتی این مدل‌ها، با دشواری‌های زیادی همراه است (دایبولت و رابرت، ۱۹۹۰). معمولاً در این حالت توزیع پسین فاقد یک فرم بسته می‌باشد. از این رو برای انجام استنباط، به کمک الگوریتم‌های زنجیر مارکوف مونت کارلوئی ($MCMC$)، از توزیع پسین نمونه‌گیری می‌شود. الگوریتم‌های گیبز و مترو پولیس-هاستینگ، از جمله مهمترین الگوریتم‌ها در زمینه نمونه‌گیری از توزیع پسین هستند. اما، برای استفاده از الگوریتم گیبز لازم است توزیع‌های شرطی کامل پارامترها در دسترس باشند. بعلاوه نشان داده شده است که وقتی این الگوریتم برای برآورد پارامترهای توزیع‌های آمیخته به کار گرفته می‌شود، همگرایی آن به مقادیر اولیّه وابسته است. از این رو اگر مقادیر اولیّه به ماکسیمم موضعی نزدیک باشند، ممکن است حتی برای تکرارهای بسیار زیاد

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

الگوریتم توانایی رهایی از جذب ماکسیمم موضعی را نداشته باشد. به همین دلیل از الگوریتم مترو پولیس–هاستینگ که در هنگام کار با توزیع‌های آمیخته فاقد معایب الگوریتم گیبز می‌باشد، برای برآورد پارامترهای مدل استفاده شده است (مارین و همکاران، ۲۰۰۵). در این الگوریتم با ساختن زنجیر مارکف $\{\theta^{(t)}\}_{t=1}^{N}$ که دارای توزیع مانای $\pi(\theta|y)$ است، مقدار $E(g(\theta)|y)$ توسط برآوردگر سازگار $\frac{1}{N}\sum_{t=1}^{N} g(\theta^{(t)})$ تخمین زده می‌شود.
فرض کنید $Y_1, ..., Y_n$ یک نمونه تصادفی از توزیع آمیخته‌ی دو مؤلفه‌ای باشد، شکل کلی الگوریتم متروپولیس – هاستینگ برای برآورد پارامترهای مدل آمیخته و نیز پارامتر h، بصورت زیر است:

آ. مقادیر اولیه‌ی $p^{(0)}$، $\theta^{(0)}$ و $h^{(0)}$ را در نظر بگیرید.
ب. مراحل زیر را تا رسیدن به توزیع مانای زنجیر مارکف $\{\theta^{(t)}\}_{t=1}^{N}$ تکرار کنید.
پ. $(\tilde\theta, \tilde p, \tilde h)$ را از توزیع پیشنهادی $q(\tilde\theta, \tilde p, \tilde h|\theta^{(t-1)}, p^{(t-1)}, h^{(t-1)})$، تولید کنید.
ت. مقدار r را محاسبه کنید،

$$r = \frac{f(y|\tilde\theta, \tilde p, \tilde h)\Pi(\tilde\theta, \tilde p, \tilde h)q(\theta^{(t-1)}, p^{(t-1)}, h^{(t-1)}|\tilde\theta, \tilde p, \tilde h)}{f(y|\theta^{(t-1)}, p^{(t-1)}, h^{(t-1)})\Pi(\theta^{(t-1)}, p^{(t-1)}, h^{(t-1)})q(\tilde\theta, \tilde p, \tilde h|\theta^{(t-1)}, p^{(t-1)}), h^{(t-1)})}.$$

ث.) مشاهده‌ی u را از توزیع $U(0,1)$ را در نظر بگیرید. اگر $r > u$ قرار دهید $(\theta^{(t)}, p^{(t)}, h^{(t)}) = (\tilde\theta, \tilde p, \tilde h)$ و در غیر اینصورت قرار دهید $(\theta^{(t)}, p^{(t)}, h^{(t)}) = (\theta^{(t-1)}, p^{(t-1)}, h^{(t-1)})$.

۵ شبیه‌سازی

به منظور ارزیابی و مقایسه‌ی تأثیر توابع تاوان (۶) و (۷) بر آزمون نسبت درستنمایی اصلاح شده، یک مطالعه‌ی شبیه‌سازی اجرا شده است. برای این منظور آمیخته‌ای از دو توزیع پواسن با پارامترهای مختلف به نحوی در نظر گرفته شده‌است، که میانگین توزیع تحت فرض‌های $\mathcal{H}_0$ و $\mathcal{H}_1$ با یکدیگر مساوی و برابر با ۵ باشد. برای این منظور از روابط

$$\begin{aligned}
E_{\mathcal{H}_0}(Y) &= Var_{\mathcal{H}_0}(Y),\\
E_{\mathcal{H}_1}(Y) &= p\lambda_1 + (1-p)\lambda_2,\\
Var_{\mathcal{H}_1}(Y) &= E_{\mathcal{H}_1}(Y^2) - E^2_{\mathcal{H}_1}(Y)\\
&= p(1-p)(\lambda_1-\lambda_2)^2 + 5,
\end{aligned}$$

استفاده شده است. بصورت مشابه آمیخته‌ای از توزیع نرمال با پارامترهای متفاوت نیز به نحوی در نظر گرفته شده است، که میانگین توزیع تحت فرض‌های $\mathcal{H}_0$ و $\mathcal{H}_1$ با یکدیگر

[DOR: 20.1001.1.17358183.1387.2.2.6.2]

مساوی و برابر با ۰ باشد. برای این منظور روابط

$$
\begin{aligned}
E_{\mathcal{H}_0}(Y) &= 0, \qquad Var_{\mathcal{H}_0}(Y) = 1, \\
E_{\mathcal{H}_1}(Y) &= p\mu_1 + (1-p)\mu_2, \\
Var_{\mathcal{H}_1}(Y) &= E_{\mathcal{H}_1}(Y^2) - E^2_{\mathcal{H}_1}(Y) \\
&= p(1-p)(\mu_1 - \mu_2)^2 + 1,
\end{aligned}
$$

مورد استفاده قرار گرفته‌اند. مدل‌های حاصل از این انتخاب‌ها در جدول ۱ نشان داده شده‌اند.

جدول ۱: توزیع‌های آمیخته‌ی پواسن و نرمال با میانگین‌های برابر تحت فرض‌های $\mathcal{H}_0$ و $\mathcal{H}_1$.

مدل	p	توزیع پواسن		توزیع نرمال	
		λ_1	λ_2	μ_1	μ_2
۱	۰/۰۵	۰/۱۲۷	۵/۲۵۶	−۲/۱۷۹	۰/۱۱۵
۲	۰/۱۰	۱/۶۴۶	۵/۳۷۳	−۱/۵۰۰	۰/۱۶۷
۳	۰/۲۵	۳/۰۶۴	۵/۶۴۵	−۰/۸۶۶	۰/۲۸۹
۴	۰/۵۰	۳/۸۸۲	۶/۱۱۸	−۰/۵۰۰	۰/۵۰۰

این چهار مدل از دو توزیع آمیخته‌ی پواسن و نرمال بصورتی در نظر گرفته می‌شود که $p = 0/5, 0/25, 0/1, 0/05$ اتخاذ شود و واریانس برای مدل‌های آمیخته ۱/۲۵ برابر واریانس تحت مدل همگن باشد و میانگین نیز برای مدل‌های آمیخته برابر با میانگین تحت مدل همگن باشد. سپس از هر یک از این مدل‌ها نمونه‌ای به حجم $n = 200$ شبیه سازی شده و فرض همگنی (۲) به کمک آزمون نسبت درستنمایی اصلاح شده و با در نظر گرفتن توابع تاوان (۶) و (۷) در سطح ۰/۰۵ آزمون شده است. نتایج حاصل که در جدول ۲ خلاصه شده‌اند، نشان می‌دهند که وقتی مدل آمیخته به سمت شناخت ناپذیری میل می‌کند، استفاده از تابع تاوان (۷) به جای تابع تاوان (۶) در آزمون نسبت درستنمایی اصلاح شده، منجر به افزایش توان آزمون و دقت برآوردها با استناد به معیار میانگین توانهای دوم خطا، می‌شود. همچنین، با استفاده از رهیافت بیزی، پارامترهای مدل آمیخته و نیز پارامتر شکل تابع تاوان برای توزیع آمیخته نرمال و پواسن برآورد شده‌اند، جداول ۳ و ۴ مقادیر میانگین توان‌های دوم خطای برآوردهای حاصل از دو رهیافت بسامدی و بیزی را نشان می‌دهند. در رهیافت بسامدی از الگوریتم EM برای برآورد پارامترهای مدل آمیخته و از تابع تاوان (۷)

جدول ۲: برآورد پارامترها، میانگین توان دوم خطا، مقدار آماره و توان آزمون نسبت درستنمایی اصلاح شده، برای توابع تاوان چن و لی.

مدل	تابع تاوان	برآورد			MSE			آماره آزمون	توان آزمون
		$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}$		(درصد)
۱	چن	۰/۷۵۰	۵/۳۸۵	۰/۱۲۱	۱/۷۹۰	۰/۰۸	۰/۰۱۶	۶۲/۵۰۷	۸۸/۲
۲	لی	۰/۳۳۳	۵/۳۰۳	۰/۰۷	۰/۴۳۹	۰/۴۳۱	۰/۰۰۳	۶۲/۶۰۰	۹۷/۴
۳	چن	۲/۹۳۹	۵/۸۱	۰/۳۴۵	۳/۱۰۳	۰/۲۸	۰/۰۷۳	۴۷/۰۱۰	۹۲/۱
۴	لی	۲/۵۷۱	۵/۶۶۴	۰/۲۶۱	۲/۱۶۶	۰/۲۰۱	۰/۰۴۸	۴۸/۹۹۵	۹۲/۲
۵	چن	۳/۷۱۷	۶/۰۱۱	۰/۴۵۲	۰/۶۷۴	۰/۲۵۲	۰/۰۴۵	۴۲/۳۰۷	۹۳/۸
۶	لی	۳/۷۱۷	۶/۰۱۱	۰/۴۵۲	۰/۶۷۴	۰/۲۵۲	۰/۰۴۵	۴۲/۳۰۷	۹۳/۹
۷	چن	۳/۹۱۹	۶/۰۸۲	۰/۵	۰/۱۴۵	۰/۱۶۵	۰/۰۰۳	۲۷/۶۹۱	۹۳/۵
۸	لی	۳/۹۱۱	۶/۰۹۰	۰/۴۹۹	۰/۱۸۲	۰/۲۱۶	۰/۰۰۵	۲۷/۵۸۸	۹۳/۴

جدول ۳: میانگین توان دوم خطای برآوردگرهای بیزی و بسامدی.

مدل پواسن	برآوردگر	پارامترهای مدل			
		$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{p}$	$\hat{h}$
۱	$\hat{\theta}_{ML}$	۱/۵۶۷۰	۰/۱۰۱۰	۰/۰۰۵۰	–
	$\hat{\theta}_B$	۰/۱۰۷۰	۰/۰۲۴۰	۰/۰۰۰۴	۰/۰۴۸۹
۲	$\hat{\theta}_{ML}$	۱/۹۱۵۰	۰/۱۸۵۰	۰/۰۳۴۰	–
	$\hat{\theta}_B$	۰/۰۹۹۰	۰/۰۲۱۰	۰/۰۰۰۴	۰/۰۵۴۰
۳	$\hat{\theta}_{ML}$	۰/۰۷۲۰	۰/۰۲۴۰	۰/۰۰۲۲	–
	$\hat{\theta}_B$	۰/۰۴۵۰	۰/۰۲۳۰	۰/۰۰۰۷	۰/۰۷۳۰
۴	$\hat{\theta}_{ML}$	۰/۰۴۰۶	۰/۰۰۳۱۰	۰	–
	$\hat{\theta}_B$	۰/۰۷۱۰	۰/۰۴۳۰	۰/۰۰۰۸	۰/۰۵۵۰

استفاده شده، در صورتی که در رهیافت بیزی از الگوریتم متروپولیس – هاستینگ برای براورد پارامترهای مدل آمیخته و نیز پارامتر h استفاده شده است. همانطور که مشاهده می‌شود، زمانی که مدل آمیخته به سمت شناخت ناپذیری میل می‌کند، استفاده از رهیافت بیزی در برآورد پارامترهای مدل به مراتب از رهیافت بسامدی مطلوب‌تر است. به منظور ارزیابی میزان

جدول ۴: میانگین توان دوم خطای برآوردگرهای بیزی و بسامدی.

مدل نرمال	برآوردگر	پارامترهای مدل			
		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{p}$	$\hat{h}$
۱	$\hat{\theta}_{ML}$	۲/۵۲۳۰	۰/۰۶۸۰	۰/۰۵۷۰	–
	$\hat{\theta}_B$	۰/۲۴۳۰	۰/۰۰۴۵	۰/۰۰۴۴	۰/۰۰۹۴
۲	$\hat{\theta}_{ML}$	۰/۱۵۰۵	۰/۰۲۱۰	۰/۰۱۶۷	–
	$\hat{\theta}_B$	۰/۰۰۳۱	۰/۰۰۰۲	۰/۰۰۲۵	۰/۱۹۴۳
۳	$\hat{\theta}_{ML}$	۰/۳۲۱۲	۰/۰۰۲۱	۰/۰۶۱۰	–
	$\hat{\theta}_B$	۰/۰۰۸۱	۰/۰۰۰۶	۰/۰۰۲۲	۰/۰۲۹۰
۴	$\hat{\theta}_{ML}$	۰/۰۰۰۲	۰/۰۳۸۰	۰	–
	$\hat{\theta}_B$	۰/۰۰۰۴	۰/۰۲۶۰	۰/۰۰۰۹	۰/۰۱۶۵

برازش مدل‌هایی که از رهیافت بیزی برآورد شده‌اند، معیار نیکویی برازش کیبش و آکائیک محاسبه و نتایج آن در جدول ۵ ارائه شده است. نکتهٔ حائز اهمیت آن است که گرچه مقادیر کوچک این معیارهای نیکویی برازش به معنی برازش دقیق مدل برآورد شده به داده‌ها نیست، اما مقادیر بزرگ آن گواهی روشن بر نامناسب بودن مدل برآورد شده برای داده‌ها است (هاسمر و لِمشاو، ۲۰۰۰). معیار نیکویی برازش کیبش و آکائیک به ترتیب از روابط

$$\begin{aligned} D &= -2\ell_n(\hat{\theta}, \hat{p}), \\ AIC &= -2\ell_n(\hat{\theta}, \hat{p}) + 2t, \end{aligned}$$

محاسبه می‌شوند، که در آن $\ell_n(\hat{\theta}, \hat{p})$ تابع لگ درستنمایی و t تعداد پارامترهای مدل می‌باشد. نتایج ارائه شده در جدول ۵ نشان می‌دهند که در هر دو مورد توزیع‌های آمیخته‌ی پواسن و نرمال با دو مؤلفه ($k = 2$) در مقابل مدل همگن ($k = 1$)، برازش به مراتب بهتری به داده‌هایی دارند که از همان توزیع‌ها شبیه‌سازی شده‌اند.

جدول ۵: مقادیر معیارهای نیکویی برازش آکائیک و کیبش.

مدل	k	توزیع نرمال		توزیع پواسن	
		AIC	$D(\theta)$	AIC	$D(\theta)$
۱	۱	۹۷۷/۰۰۰	۹۷۵/۹۰۰	۱۴۴۸/۸۰۰	۱۴۴۶/۸۰۰
	۲	۹۶۴/۸۰۰	۹۵۶/۸۰۰	۱۴۰۴/۰۸۰	۱۳۹۶/۰۸۰
۴	۱	۹۳۹/۰۸۴	۹۳۴/۰۷۱	۱۴۰۰/۲۹۵	۱۳۹۸/۲۹۵
	۲	۹۳۵/۴۱۸	۹۲۷/۴۸۳	۱۳۰۰/۰۰۵	۱۳۹۲/۰۰۵

۶ بحث و نتیجه‌گیری

کارایی آزمون نسبت درستنمایی اصلاح شده به شدت تحت تأثیر شکل تابع تاوان انتخابی است. از طرفی هیچ دلیلی مبنی بر بهینه بودن هیچ یک از این توابع تاوانی که در منابع مختلف ارائه شده‌اند، وجود ندارد. استفاده از رهیافت بیزی هم در برآورد پارامترهای مدل آمیخته و هم برای تعیین شکل بهینه تابع تاوان خصوصاً در شرایطی که مدل آمیخته به سمت شناخت ناپذیری میل می‌کند، به مراتب از رهیافت بسامدی مطلوب‌تر است. آزمون همگنی توزیعهای آمیخته، عمدتاً به منظور تشخیص آمیخته بودن یا نبودن توزیع جامعه صورت می‌پذیرد. تعیین تعداد مؤلفه‌های توزیع آمیخته در مرحلهٔ بعدی قرار دارد.

مراجع

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, *In Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki (Eds). Budapest: Akademiai Kiado, 261-281.

Chen, H., Chen, J. and Kalbfleish, D. (2001), The Likelihood Ratio Test for Homogeneity in Finite Mixture Models, *Journal of the Royal Statistical Society*, **63**, 19-29.

Chen, J., (1998), Penalized Likelihood Ratio Test for Finite Mixture Models with Multinomial Observations, *Canadian Journal of Statistics*, **26**, 583-599.

Chen, J. and Kalbfleisch, J. D. (1996), Penalized Minimum-Distance Estimates in Finite Mixture Models, *Canadian Journal of Statistics*, **24**, 167-175.

Chen, J. and Khalili, A. (2006), Order Selection in Finite Mixture Models, *Working Paper 2006-03*, Department of Statistics and Actuarial Science, University of Waterloo.

Diebolt, J. and Robert, C. (1990), Bayesian Estimation of Finite Mixture Distribution, Part i: Theoretical Aspects. *Technical Report 110*, LSTA, Universite Paris VI, Paris.

Hartigan, J. A. (1985), A Failure of Likelihood Asymptotics for Normal Mixtures, *Proceedings of conference in Honor of J. Neyman and Kiefer*, Volume 2, eds L. LeCam and R. A. Olshen, 807-810.

Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression*, Wiley, Inc., New York.

Li, P. Chen, J. and Marrriott, P. (2008), Non-Finite Fisher Information and Homogeneity: The EM Approach, *Biometrika*. **96**, 411-426.

Marin, J.-M., Mengerson, K. and Robert, C. (2005), Bayesian Modelling and Inference on Mixture of Distributions, *Handbook of Statistics*, **25**, 459-507, DOI: 10.1016/S0169-7161(05)25016-2.

Richardson, S. and Green, P. J. (1997), On Bayesian Analysis of Mixtures with an Unknown Number of Components, *Journal of the Royal Statistical Society*, B, **59**, 731-792.