تعیین تعداد مؤلفه‌ها در توزیع آمیخته با مؤلفه‌های تی-نرمال‌چوله

محمد بهرامی[۱]، فهیمه طورانی فرانی [۲]

تاریخ دریافت: ۱۳۹۱/۸/۲۸
تاریخ پذیرش: ۱۳۹۶/۴/۲۵

چکیده:

تعیین تعداد مؤلفه‌ها در یک توزیع آمیخته، مسئله‌ای دشوار و حائز اهمیت است. برای تعیین تعداد بهینه مؤلفه‌ها در توزیع‌های آمیخته، روش‌های مختلفی وجود دارد که در این مقاله به ذکر چند مورد از آنها خواهیم پرداخت. روش اول که تحت عنوان الگوریتم greedy EM بیان شده، بر اساس الگوریتمی است که طی هر مرحله آن مؤلفه‌ای جدید به مدل اضافه می‌شود و این روند تا زمانیکه منجر به تعیین تعداد بهینه مؤلفه‌ها در توزیع آمیخته شود، ادامه می‌یابد. روش دوم بر اساس ماکسیمم آنتروپی ادغام در تکرار زیرکلاس‌های روی هم افتاده تا زمانی است که در نتیجهٔ ادغام این مؤلفه‌ها، توزیع آمیخته مورد بررسی دارای یک مؤلفه شود. این روش با عنوان ادغام آمیختگی شرح داده شده است و روش سوم نیز توسط تعریف متغیرهای نشانگر، به‌صورت ناپارامتری تعداد مؤلفه‌های توزیع آمیخته را تعیین می‌کند. شایان ذکر است که مؤلفه‌های توزیع آمیخته مورد نظر در این مقاله توزیع تی-نرمال‌چوله در نظر گرفته شده‌اند.

واژه‌های کلیدی: توزیع آمیخته متناهی، تعداد مؤلفه‌ها، الگوریتم greedy EM، آنتروپی، ادغام آمیختگی، توزیع تی-نرمال‌چوله.

۱ معرفی توزیع آمیخته

فرض کنید جامعه آماری P شامل k ($k \geq ۲$) زیر جامعه $P_۱$ و $P_۲$ و $\cdots$ و P_k باشد. هدف، بررسی خصوصیتی در این جامعه است که متغیر تصادفی X نشان‌دهندهٔ اندازه آن خصوصیت است. اگر $F_۱$ و $F_۲$ و $\cdots$ و F_k تابع توزیع متغیر تصادفی X در زیر جامعه‌های $P_۱$ و $P_۲$ و $\cdots$ و P_k باشند، آنگاه تابع توزیع متغیر تصادفی X با تکیه‌گاه S در جامعه P به‌صورت

(۱)

$$F(x) = \pi_۱ F_۱(x) + \pi_۲ F_۲(x) + \cdots + \pi_k F_k(x), \quad x \in S$$

است، که در آن π_i احتمال تعلق X به زیرجامعه P_i و دارای توزیع F_i است. برای $i = ۱, ۲, ..., k$، $۰ < \pi_i < ۱$ و $\sum_{i=۱}^{k} \pi_i = ۱$. مقادیر π_i را نسبت‌های آمیختگی می‌نامند. مدل **(؟؟)** یک توزیع آمیخته متناهی با k مؤلفه بر حسب تابع توزیع است. چنانچه $f_۱$ و $f_۲$ و $\cdots$ و f_k به ترتیب توابع چگالی متغیر تصادفی X در زیرجامعه‌های $P_۱$ و $P_۲$ و $\cdots$ و P_k باشند، آنگاه مدل آمیخته بر حسب توابع چگالی عبارت‌اند از:

$$f(x) = \pi_۱ f_۱(x) + \pi_۲ f_۲(x) + \cdots + \pi_k f_k(x)$$
$$= \sum_{i=۱}^{k} \pi_i f_i(x); \quad x \in S. \qquad (۲)$$

در حالت کلّی لزومی ندارد که مؤلفه‌های $f_۱$ و $f_۲$ و $\cdots$ و f_k متعلق به یک خانواده از توابع چگالی باشند، اما در اغلب موارد، این اتفاق رخ می‌دهد. توابع چگالی f_i برای $i = ۱, ۲, \ldots, k$ می‌توانند شامل بردار پارامتر $\boldsymbol{\theta}_i = (\theta_{i۱}, \theta_{i۲}, ..., \theta_{ih})^T$ باشند. بنا بر این یک چگالی آمیخته‌متناهی پارامتری با k مؤلفه به شکل کلّی

$$f(x|\boldsymbol{\theta}) = \sum_{i=۱}^{k} \pi_i f(x|\theta_i); \quad \boldsymbol{\theta} \in \Theta,\ x \in S \qquad (۳)$$

$\boldsymbol{\theta} = (\pi_۱, \pi_۲, ..., \pi_k, \boldsymbol{\theta}_۱^T, \boldsymbol{\theta}_۲^T, ..., \boldsymbol{\theta}_k^T)^T$ است. را بردار پارامترهای مدل آمیخته می‌نامیم که در فضای پارامتر Θ تغییر

[۱] عضو هیئت علمی گروه آمار، دانشگاه اصفهان، ایران
[۲] دانشجوی دکتری، دانشگاه اصفهان، ایران

می‌کند. به عنوان مثال تابع چگالی توزیع آمیخته با مؤلفه‌های تی-نرمال‌چوله به‌صورت

$$f(y|\boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i g(y|\mu_i, \sigma_i^2, \alpha_i, v_i)$$

است، به‌طوری که

$$g(y|\mu, \sigma^2, \alpha, v) = 2t(y|\mu, \sigma^2, v)\Phi\left(\alpha\frac{y-\mu}{\sigma}\right)$$

و نمادهای t و Φ به ترتیب تابع چگالی احتمال توزیع تی و تابع توزیع تجمعی نرمال هستند و μ و α نیز به ترتیب پارامترهای مکان و شکل بوده و متعلق به اعداد حقیقی $\mathbb{R}$ هستند همچنین σ و v به ترتیب پارامترهای مقیاس و درجهٔ آزادی هستند که متعلق به بازه $(0, +\infty)$ هستند. بردار پارامترهای توزیع، θ، به‌صورت $\boldsymbol{\theta} = (\pi_1, \pi_2, ..., \pi_k, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, ..., \boldsymbol{\theta}_k^T)^T$ بوده که در آن $\boldsymbol{\theta}_i = (\mu_i, \sigma_i^2, \alpha_i, v_i)$، $0 < \pi_i < 1$ و $\sum_{i=1}^{k} \pi_i = 1$. پارامتر k، تعداد مؤلفه‌ها در یک توزیع آمیخته، از اهمیت زیادی برخوردار است. زیرا اگر k را بسیار بزرگ در نظر بگیریم، ممکن است میانگین برخی از زیرجامعه‌ها بسیار شبیه به یکدیگر باشند و یا π_j (احتمال تعلق به زیرجامعه jام) متناظر با برخی از آنها بسیار کوچک باشد و چنانچه k را مقداری کوچک در نظر بگیریم، آنگاه ممکن است μ_jها نماینده خوبی برای هر گروه نباشند و یا به‌عبارت دیگر زیرجامعه‌ها دارای واریانس بزرگی باشند. نکته دیگر که اهمیت معلوم بودن تعداد مؤلفه‌های یک توزیع آمیخته را روشن‌تر می‌کند، تعیین تعداد پارامترهای مدل است. زیرا چنانچه k مقداری معلوم باشد، تعداد پارامترهای توزیع نیز معلوم است. به عنوان مثال یک توزیع آمیخته شامل k مؤلفه نرمال با میانگین و واریانس متفاوت، دارای $2k + (k-1)$ پارامتر است. بنا بر این با توجه به اهمیت تعداد مؤلفه‌ها در یک توزیع آمیخته، روش‌های زیادی برای تعیین تعداد بهینه این مؤلفه‌ها بیان شده است که در این مقاله به بررسی سه مورد از آنها می‌پردازیم. الگوریتم gEM برای حالت خاص توزیع آمیخته با مؤلفه‌های تی-نرمال‌چوله بیان شده، ولی روش کلّی آن را می‌توان برای هر توزیع آمیخته دلخواه به کار برد. دو روش دیگر در حالت کلّی و برای هر توزیع آمیخته بیان شده‌اند.

۲ روش الگوریتم gEM

روش gEM[3] اولین بار توسط ولاسیس و لیکاس [۴] معرفی شد. در این روش کمترین مقدار مؤلفه برای یک توزیع آمیخته در نظر گرفته می‌شود و سپس طی چند مرحله با اضافه کردن مؤلفه جدید به آن، مدل مناسب برای توصیف داده‌ها و در نتیجه تعداد بهینه مؤلفه‌ها برای توزیع آمیخته اولیه تعیین می‌شود.

ابتدا توسط الگوریتم امید ریاضی ماکسیمم‌سازی (EM)[4] و یا امید ریاضی ماکسیمم‌سازی شرطی (EMC)،[5] پارامترهای توزیع آمیخته اولیه با کمترین مؤلفه (k مؤلفه) براورد می‌شود. توجه کنید که بردار پارامترهای هر مؤلفه به‌صورت

$$\boldsymbol{\theta}_i = (\mu_i, \sigma_i^2, \alpha_i, v_i)^T$$

است.

فرض کنید مؤلفه جدید $g(y_j|\boldsymbol{\theta}_{k+1})$ به توزیع آمیخته k مؤلفه‌ای اولیه، $f(y_j|\Theta_k)$، اضافه شده است، آنگاه توزیع آمیخته جدید به‌صورت

$$f(y_j|\Theta_{k+1}) = (1-a)f(y_j|\Theta_k) + ag(y_j|\boldsymbol{\theta}_{k+1}) \quad (۴)$$

است، که در آن $0 < a < 1$ و $\Theta_{k+1} = (\Theta_k, a, \Theta_{k+1}^T)^T$ و $\Theta_{k+1} = (\mu_{k+1}, \sigma_{k+1}^2, \alpha_{k+1}, v_{k+1})^T$ است. a و $1-a$ به ترتیب نسبت‌های آمیختگی برای توزیع آمیخته اولیه و مؤلفه جدید هستند. سپس پارامترهای جدید a و Θ_{k+1} توسط ماکسیمم سازی لگاریتم تابع درست‌نمایی

$$\begin{aligned}\mathcal{L}_{k+1} &= \sum_{j=1}^{n} \log f(y_j|\Theta_{k+1}) \\ &= \sum_{j=1}^{n} \log\left[(1-a)f(y_j|\hat{\Theta}_k) + ag(y_j|\Theta_{k+1})\right] \quad (۵)\end{aligned}$$

براورد می‌شوند. با توجه به تابع درست‌نمایی فوق مشاهده می‌شود که پارامترهای توزیع آمیخته اولیه در طول ماکسیمم

[3] greedy EM

[4] Expectation and Maximization

[5] Conditional Expectation and Maximization

سازی $\mathcal{L}_{k+1}$ ثابت باقی می‌مانند.

نکته حائز اهمیت در این روش تبدیل یک توزیع آمیخته با $k+1$ مؤلفه به توزیعی آمیخته با دو مؤلفه است که منجر به سهولت در محاسبات براوردیابی می‌شود. برای براورد بردار پارامترهای Θ_{k+1}، پارامترهای توزیع آمیخته جدید با دو مؤلفه، استفاده از الگوریتم EM مناسب است. اما با توجه به این‌که پارامترهای Θ_k در طول فرایند براوردیابی به عنوان یک مقدار ثابت عمل می‌کنند، این الگوریتم بخشی از تابع درست‌نمایی را ماکسیمم می‌کند و از این رو الگوریتم EM جزئی نامیده می‌شود. محاسبات متناظر با این براوردها مشابه با محاسبات متناظر با براورد پارامترهای توزیع آمیخته اولیه با k مؤلفه است. گام‌های الگوریتم EM جزئی به شرح زیر است.

- گام امید ریاضی جزئی: در این مرحله امید ریاضی شرطی متغیرهای پنهان به شرط مشاهدات در m امین تکرار محاسبه می‌شود. در اینصورت

$$\begin{aligned}
\hat{z}_j^{(m)} &= \frac{\hat{a}^{(m)} g\left(y_j|\hat{\Theta}^{(m)}\right)}{\left(1-\hat{a}^{(m)}\right) f\left(y_j|\hat{\Theta}_k^{(m)}\right) + \hat{a}^{(m)} g\left(y_j|\hat{\Theta}^{(m)}\right)} \\
\hat{t}_j^{(m)} &= \frac{\hat{v}^{(m)}+1}{\hat{v}^{(m)}+\hat{u}_j^{2(m)}} \\
\hat{\gamma}_{1j}^{(m)} &= \hat{\alpha}^{(m)}\hat{u}_j^{(m)} + \frac{\varphi(\hat{\alpha}^{(m)}\hat{u}_j^{(m)})}{\Phi(\hat{\alpha}^{(m)}\hat{u}_j^{(m)})} \\
\hat{k}_j^{(m)} &= DG\left(\frac{\hat{v}^{(m)}+1}{2}\right) - \log\left(\frac{\hat{v}^{(m)}+\hat{u}_j^{2(m)}}{2}\right) \qquad (۶)
\end{aligned}$$

است، به‌طوری که $\hat{u}_j^{(m)} = \left(y_j - \hat{\mu}^{(m)}\right)/\hat{\sigma}^{(m)}$ است.

- گام ماکسیمم سازی جزئی: پارامترهای جدید موجود در (a, Θ_{k+1})، توسط ماکسیمم سازی امید ریاضی شرطی تابع درست‌نمایی (۵) به شرط مشاهدات، براورد می‌شوند. براوردهای حاصل به‌صورت

$$\begin{aligned}
\hat{a}^{(m+1)} &= \frac{\sum_{j=1}^n \hat{z}_j^{(m)}}{n} \\
\hat{\mu}_{k+1}^{(m+1)} &= \frac{\hat{b}_1^{(m)} + \hat{\alpha}^{2(m)}\hat{b}_2^{(m)} - \hat{\sigma}^{(m)}\hat{\alpha}^{(m)}\hat{b}_3^{(m)}}{\sum_{j=1}^n \hat{z}_j^{(m)}\hat{t}_j^{(m)} + \hat{\alpha}^{2(m)}\sum_{j=1}^n \hat{z}_j^{(m)}} \\
\hat{\sigma}_{k+1}^{2(m+1)} &= \frac{\sum_{j=1}^n \hat{z}_j^{(m)}\hat{t}_j^{(m)}\left(y_j - \hat{\mu}^{(m+1)}\right)^2}{\sum_{j=1}^n \hat{z}_j^{(m)}} \\
\hat{\alpha}_{k+1}^{(m+1)} &= \frac{\sum_{j=1}^n \hat{z}_j^{(m)}\hat{\gamma}_{1j}^{(m)}\hat{\mu}_j^{(m+1)}}{\sum_{j=1}^n \hat{z}_j^{(m)}\hat{\mu}_j^{2(m+1)}} \\
\hat{v}_{k+1}^{(m+1)} &= \operatorname{argmax}\left\{\frac{v}{2}\log\frac{v}{2} - \log\Gamma\left(\frac{v}{2}\right) + \frac{v}{2}\hat{b}_4^{(m)}\right\} \qquad (۷)
\end{aligned}$$

است که در آن

$$\begin{aligned}
\hat{u}_j^{(m+1)} &= \left(y_j - \hat{\mu}^{(m+1)}\right)/\hat{\sigma}^{(m+1)} \\
\hat{b}_1^{(m+1)} &= \sum_{j=1}^n \hat{z}_j^{(m)}\hat{t}_j^{(m)} y_j \\
\hat{b}_2^{(m+1)} &= \sum_{j=1}^n \hat{z}_j^{(m)} y_j \\
\hat{b}_3^{(m+1)} &= \sum_{j=1}^n \hat{z}_j^{(m)}\hat{\gamma}_{1j}^{(m)} \\
\hat{b}_4^{(m+1)} &= \sum_{j=1}^n \hat{z}_j^{(m)}\left(\hat{k}_j^{(m)}\hat{t}_j^{(m)}\right) / \sum_{j=1}^n \hat{z}_j^{(m)}
\end{aligned}$$

هستند. گام‌های الگوریتم EM جزئی را تا رسیدن به همگرایی به کار می‌بریم. به عنوان مثال تا وقتی که

$$\left|\frac{\hat{\mathcal{L}}^{(m)_{k+1}}}{\hat{\mathcal{L}}^{(m-1)_{k+1}}} - 1\right| < 10^{-6}$$

گام‌های الگوریتم را تکرار می‌کنیم.

اگرچه الگوریتم EM جزئی روشی کارآمد در براوردیابی پارامترهای توزیع آمیخته حاصل است، اما حساسیت این روش به مقادیر اولیه a و μ_{k+1} غیرقابل اغماض است. بنا بر این برای یافتن مقادیر اولیه مناسب این پارامترها تدبیر جست و جوی کلی را به کار می‌بریم. اساس این روش به‌منظور یافتن یک مقدار اولیه مناسب و همچنین تسهیل جست و جوی کامل در فضای پارامتر، جایگزینی تابع درست‌نمایی (۵) با بسط تیلور در نقطه $a = a.$ و سپس یافتن مقدار بهینه پارامترهای Θ_{k+1} است. به‌ویژه اگر از بسط تیلور مرتبه دو استفاده کنیم، تقریب زیر را مشاهده خواهیم کرد.

$$\hat{\mathcal{L}}_{k+1} = \mathcal{L}_{k+1}(a.) + \frac{\left|\dot{\mathcal{L}}_{k+1}(a.)\right|^2}{2\ddot{\mathcal{L}}_{k+1}(a.)} \qquad (۸)$$

که در آن $\dot{\mathcal{L}}_{k+1}(a.)$ و $\ddot{\mathcal{L}}_{k+1}(a.)$ مشتق‌های مرتبه اول و دوم $\mathcal{L}_{k+1}(a.)$ حول نقطه $a = a.$ است.

با بررسی رابطه فوق می‌توان نشان داد که ماکسیمم $\mathcal{L}_{k+1}$ حول نقطه$a. = ۰/۵$ به‌صورت

$$\hat{\mathcal{L}}_{k+1} = \sum_{j=1}^{n} \log\left(\frac{f\left(y_j|\hat{\Theta}_k\right) + g\left(y_j|\Theta_{k+1}\right)}{2}\right) + \frac{\left[\sum_{j=1}^{n}\delta_j(\Theta_{k+1})\right]^2}{2\sum_{j=1}^{n}\delta_j^2(\Theta_{k+1})} \quad (۹)$$

است. به‌طوری که

$$\delta_j(\Theta_{k+1}) = \frac{f\left(y_j|\hat{\Theta}_k\right) - g\left(y_j|\Theta_{k+1}\right)}{f\left(y_j|\hat{\Theta}_k\right) + g\left(y_j|\Theta_{k+1}\right)}.$$

بنا بر این مقدار مناسب نسبت آمیختگی a عبارت‌اند از

$$\hat{a} = \frac{1}{2}\left(1 - \frac{\sum_{j=1}^{n}\delta_j(\Theta_{k+1})}{2\sum_{j=1}^{n}\delta_j^2(\Theta_{k+1})}\right). \quad (۱۰)$$

چنانچه مقدار $\hat{a}$ خارج از بازه $(۰,۱)$ قرار گیرد، می‌توان برای $k=1$، مقدار $\hat{a}$ را برابر با ۰/۵ و برای $k \geq 2$، $\hat{a}$ را برابر با $\frac{2}{k+1}$ در نظر گرفت.

هو و همکاران [۳] ریشه پنجم نیمی از واریانس نمونه $\left(\sqrt[5]{s_y^2/2}\right)$ را یک انتخاب مناسب برای $\hat{\sigma}_{k+1}^{2(0)}$ و همچنین مقادیر ثابت صفر و ده را برای $\hat{\alpha}_{k+1}^{(0)}$ و $\hat{v}_{k+1}^{(0)}$ در نظر گرفتند. همچنین چندک‌های ۵–ام، ۱۰–ام و...و ۹۵–ام مشاهدات $\mathbf{y}$ را یافته و سپس مقداری از آنها که رابطه (؟؟) را ماکسیمم می‌کند را به عنوان $\hat{\mu}_{k+1}^{(0)}$ پیشنهاد کردند.

در حالت کلی، گام‌های الگوریتم gEM به شرح زیر است.

۱. با مقدار$k=1$ شروع کرده و پارامترهای تک مؤلفهٔ توزیع آمیخته اولیه را براورد می‌کنیم. چنانچه الگوریتم در گام اول با توزیع آمیخته بیشتر از یک مؤلفه $(k>1)$ آغاز شود، پارامترهای Θ_k با استفاده از الگوریتم EM قابل براورد هستند.

۲. توزیع آمیخته جدید با دو مؤلفه را به‌صورت رابطه (؟؟) می‌سازیم.

۳. مقادیر اولیه مناسب برای a و Θ_{k+1} را توسط روش جست و جوی کلی انتخاب می‌کنیم.

۴. پارامترهای a و Θ_{k+1} را با استفاده از الگوریتم EM جزئی براورد می‌کنیم.

۵. اگر $\hat{\mathcal{L}}_{k+1} \leq \hat{\mathcal{L}}_{k+p}$ باشد (که در آن $p>0$ یک عبارت جریمه است)، الگوریتم gEM پایان می‌یابد. در غیر اینصورت یک مؤلفه جدید در نظر می‌گیریم، به مرحله (۲) بازگشته و k را برابر با $k+1$ قرار می‌دهیم.

۳ روش ادغام آمیختگی

بادری و همکاران [۱] روش دیگری جهت تعیین تعداد بهینه مؤلفه‌ها در توزیع‌های آمیخته، تحت عنوان ادغام آمیختگی [۶] معرفی کردند. چنانچه داده‌ها به علت داشتن توزیعی با شکل غیر معمول و زیرجامعه‌های غیر قابل تفکیک [۷] و یا دارای ویژگی‌های خاصی باشند که منجر به روی هم افتادن مؤلفه‌ها و در نتیجه بیش‌براورد تعداد آنها شود، روش ادغام آمیختگی یک براورد مقاوم [۸] برای تعداد مؤلفه‌های توزیع ارائه می‌کند. در حالی‌که الگوریتم gEM یک روش مناسب برای انتخاب تعداد مؤلفه‌های یک توزیع آمیخته تحت شرایط متعارف (مانند محدب بودن مؤلفه‌ها) فراهم می‌کند. نخست روش ادغام آمیختگی برای توزیع‌های آمیخته با مؤلفه‌های متقارن، طراحی شد. اما به‌منظور رفع نیاز به مؤلفه‌هایی که در حالت متقارن توضیح دهنده چولگی توزیع واقعی داده‌ها هستند، تکنیک ادغام آمیختگی برای توزیع‌های آمیخته با مؤلفهٔ نامتقارن، گسترش یافت. به عنوان مثال چنانچه k مؤلفه‌ی یک توزیع آمیخته دارای توزیع نرمال‌چوله باشد و بخواهیم این توزیع آمیخته را توسط یک توزیع آمیخته دیگر با مؤلفه‌های نرمال بازنویسی کنیم، در این‌صورت توزیع آمیخته جدید دارای $2k$ مؤلفهٔ نرمال است. در حقیقت هر توزیع نرمال‌چوله را می‌توان توسط یک توزیع آمیخته با دو مؤلفهٔ نرمال تقریب زد که یکی از این مؤلفه‌ها در واقع توضیح دهندهٔ چولگی

[۶] Merging mixture

[۷] low separation

[۸] Robust

توزیع هستند. ماکسیمم آنتروپی ادغام در تکرار زیرکلاس‌های روی هم افتاده تا زمانی‌که در نتیجهٔ این ادغام‌ها، توزیع آمیخته مورد بررسی دارای یک مؤلفه شود، نقش اساسی را در این روش ایفا می‌کند. روش ادغام آمیختگی طی گام‌های زیر اجرا می‌شود.

۱. ابتدا بیشترین تعداد مؤلفه برای توزیع آمیخته در نظر گرفته می‌شود و میانگین آنتروپی آن توسط فرمول

$$Ent(k) = -\sum_{j=1}^{n}\sum_{i=1}^{k} \hat{z}_{ij}\log \hat{z}_{ij} \geq 0 \qquad (11)$$

محاسبه می‌شود. که در آن $\hat{z}_{ij}$ نشان‌دهنده احتمال پسین پیشامد $(\Theta_k = \hat{\Theta}_k)$ است.

۲. برای تمام زوج زیرکلاسهای ممکن $(\ell, \acute{\ell})$ دو زیر کلاس ℓ و $\acute{\ell}$ را با یکدیگر ترکیب کرده تا ضابطه زیر ماکسیمم شود.

$$\Bigg(-\sum_{i=1}^{n}\{\hat{z}_{i\ell}\log \hat{z}_{i\ell} + \hat{z}_{i\acute{\ell}}\log \hat{z}_{i\acute{\ell}}\} + \sum_{i=1}^{n}(\hat{z}_{i\ell}+\hat{z}_{i\acute{\ell}})\log(\hat{z}_{i\ell}+\hat{z}_{i\acute{\ell}})\Bigg)$$

۳. سپس دو زیرکلاس مذکور با یکدیگر ادغام می‌شود.

۴. آنتروپی حاصل پس از ادغام به‌صورت

$$Ent(k-1) = -\sum_{j=1}^{n}\left\{\sum_{i\neq \ell,\acute{\ell}} \hat{z}_{ij}\log \hat{z}_{ij} + \hat{z}_{i,\ell\cup\acute{\ell}}\log \hat{z}_{i,\ell\cup\acute{\ell}}\right\} \qquad (12)$$

محاسبه می‌شود، که در آن $\hat{z}_{i,\ell\cup\acute{\ell}} = \hat{z}_{i,\ell} + \hat{z}_{i,\acute{\ell}}$ احتمال پسین متناظر با زیر کلاس حاصل از ادغام یعنی زیرکلاس $\ell \cup \acute{\ell}$ است.

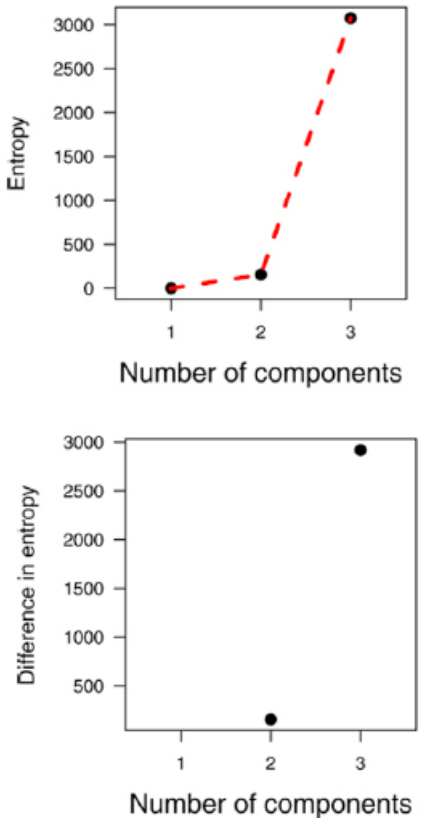
۵. مقدار $\hat{z}_j$ که شامل احتمال پسین ادغام شده و ادغام نشده است، به‌هنگام می‌شود.

۶. مقدار k برابر با $k-1$ در نظر گرفته می‌شود، به مرحله ۲ بازگشته و این کار تا $k=1$ تکرار می‌شود.

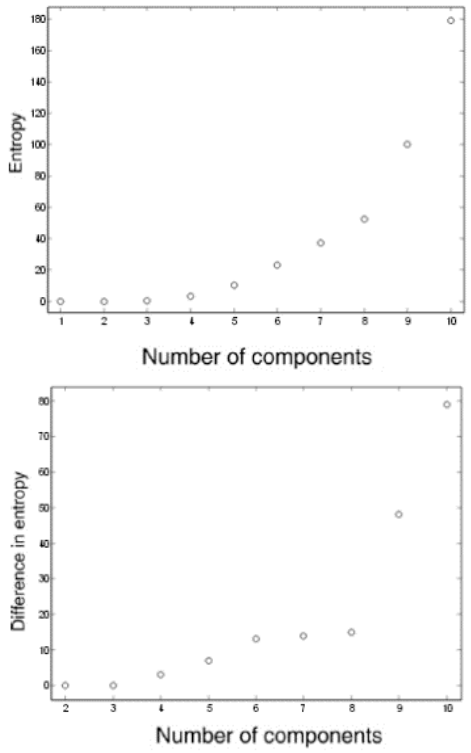
۷. در نهایت تعداد بهینه مؤلفه‌ها در توزیع آمیخته مورد بررسی برابر با نقطه‌ای است که یک پرش ناگهانی در نمودار آنتروپی زیرکلاسها در مقابل تعداد آنها دارد و یا نقطه ماکسیمم نمودار تعداد زیرکلاس‌ها در مقابل تفاضل آنتروپی آنها پس از ادغام است. به عنوان مثال نمودارهای موجود در شکل۱ (الف) بیان‌کنندهٔ تعداد دو مؤلفه و نمودارهای شکل۱ (ب) بیان‌کنندهٔ تعداد ده مؤلفه برای داده‌های متناظرشان هستند.

۴ روش ناپارامتری تعیین تعداد مؤلفه‌های توزیع آمیخته

در این بخش برای محاسبه براورد تعداد مؤلفه‌های مدل آمیخته روشی متفاوت از دو روش مطرح شده در بخش‌های قبل ارائه می‌کنیم. در حقیقت این روش، تکنیکی ناپارامتری است که توسط تعریف متغیرهای نشانگر تعداد بهینه مؤلفه‌های مدل را به شرح زیر است.



(الف)



(ب)

شکل۱. نمودار تعداد مؤلفه‌های توزیع آمیخته در مقابل آنتروپی توزیع و تفاضل آنتروپی‌های متناظر

ابتدا بیشترین تعداد مؤلفه (M) را برای توزیع آمیخته مورد نظر، فرض می‌کنیم. این تعداد را می‌توان با توجه به (نمودار فراوانی) بافت‌نگاشت داده‌ها تخمین زد. سپس جامعه را به M زیر جامعه تفکیک‌کرده و بررسی می‌کنیم که هر یک از زیرجامعه‌ها چه تعداد داده را تحت پوشش خود قرار می‌دهند و در نهایت تعداد زیرجامعه‌هایی که شامل هیچ داده‌ای نهستند (r) را شناسایی کرده و از زیر جامعه‌ها حذف می‌کنیم. بنا بر این تعداد مؤلفه‌های براورد شده توسط این روش برابر با $M-r$ است. جدول ۱ چگونگی این فرایند را بیان می‌کند. زمانی‌که مقدار M مشخص شد، اگر مشاهده i–ام متعلق به زیرجامعه

j–ام باشد، در خانه سطر i–ام و ستون j–ام عدد یک و در بقیه خانه‌های آن سطر عدد صفر را قرار می‌دهیم. این عمل را برای $i=۱,۲,...,N$ و به‌ازای $j=۱,۲,...,M$ تکرار می‌کنیم. در نهایت اعداد هر ستون را جمع کرده و سپس مشاهده می‌شود که تعداد ستون‌هایی که مجموع متناظر با آنها بزرگ‌تر از یک باشد، تعداد مؤلفه‌های مدل را نشان می‌دهند.

جدول۱. انتخاب تعداد مؤلفه‌های مدل

$\longrightarrow$ / $\downarrow$	۱	۲	$\cdots$	M
$y_۱$				
$y_۲$				
$\cdots$				
y_N				

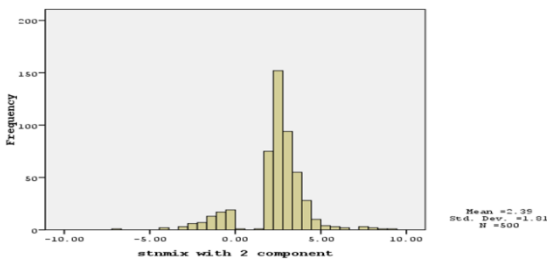
با به کارگیری برنامه زیر در نرم‌افزار Open-Bugs می‌توان تعداد مؤلفه‌های یک توزیع آمیخته را توسط روش ناپارامتری فوق تعیین کرد.

blueModel

```
{for (i in 1:n) {
for ( j in 1:M) { Memb[i,j] <-
equals(y[i],j)}}
Mstar<- sum(Fmemb[]) ;  for (j in 1:M){ Fmemb[j] <-
step(sum(Memb[,j])-1)}}
```

۵ بررسی تعداد مؤلفه‌های داده‌های شبیه‌سازی شده از توزیع آمیخته

در این بخش ابتدا یک نمونه ۵۰۰ تایی از توزیع آمیخته تی-نرمال‌چوله به گونه‌ای تولید می‌کنیم که دارای دو زیرجامعه StN با پارامترهای $(۴,-۳,۱,۰)$ و $(۴,۵,۱,۲)$ باشند. سپس با استفاده از روش ناپارامتری ارائه شده در این مقاله، تعداد مؤلفه‌های این مجموعه داده‌ها را براورد می‌کنیم. نمودار فراوانی داده‌های تولیدی به‌صورت زیر است.



شکل۲. نمودار فراوانی داده‌های شبیه‌سازی شده از توزیع آمیخته با دو مؤلفه StN

با توجه به این نمودار، واضح است که داده‌ها دارای دو زیر جامعه هستند، ولی با این وجود عدد M ، یعنی بیشترین تعداد مؤلفه در این توزیع آمیخته را با توجه به نمودار عدد ۴ در نظر گرفته و با اجرای این برنامه تعداد بهینه مؤلفه‌ها تعیین می‌شود.

جدول ۲ حاصل از اجرای برنامه مذکور برای داده‌های شبیه‌سازی شده است. با توجه به نتایج جدول ۲، تعداد مؤلفه‌ها، ۲ براورد شده است.

جدول ۲. جدول حاصل

از خروجی نرم‌افزار اپن باگز، شامل براورد تعداد مؤلفه‌های توزیع آمیخته

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
Mstar	2.0	0.0	2.236E-12	2.0	2.0	2.0	2001	2000

۶ نتیجه‌گیری

با توجه به اهمیت پارامتر تعداد مؤلفه‌های یک توزیع آمیخته، در این مقاله سه روش متفاوت برای تعیین تعداد پارامترهای توزیع آمیخته بیان شده است. چناچه تنها در مورد آمیخته بودن توزیع داده‌های مورد بررسی اطلاع داشته باشیم و از توزیع مؤلفه‌های آن اطلاعی نداشته باشیم، روش ناپارامتری بیان شده در بخش ۴، استفاده می‌شود و در صورتی‌که الگوریتم EM از نظر اجرایی و تحلیلی در توزیع آمیخته مورد بررسی ساده باشد، الگوریتم gEM و همچنین اگر توزیع داده‌ها دارای شکلّی نامعمول با زیرجامعه‌های غیر قابل تفکیک باشد روش ادغام آمیختگی به عنوان روشی مناسب توصیه می‌شود.

مراجع

[1] Baudry, J. P., Raftery, A. F., Celeux, G., Lo, K. and Gottardo, R. (2010). Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics* **9**, 332-353.

[2] Congdon, P. (2006). Bayesian Statistical Modeling. (Second Edition). *Wiley Series in probablityand statistics.*

[3] Ho, H. J., Lin, T.I., Chang, H. H., Haase, S. B., Huang, S. and Pyne, S. (2011). Parametric Modeling of Cellular State Transitions as Measured with Flow Cytometry, *frome first IEEE international Conference on Computational Advances in Bio and medical Sciences.*

[4] Vlassis, N. and Likas, A. (2002). A Greedy EM Algorithm for Gaussian Mixture Learning. *Neural Processing Letters* **15**, 77–87.