سیستم های فازی و کاربردها
شاپای چاپی: ۲۷۱۷ـ۴۴۰۹
سال هفتم، شماره اول (بهار و تابستان ۱۴۰۳)
شاپای الکترونیکی: ۲۷۱۷ـ۳۹۸۴
صص. ۱۳ تا ۳۲
DOI:10.22034/JFSA.2024.458851.1206

یک روش کلاس‌بندی فازی برای رگرسیون غیرخطی آمیخته

فرزانه هاشمی

گروه آمار، دانشکده علوم ریاضی، دانشگاه کاشان، کاشان، ایران

تاریخ دریافت: ۱۴۰۳/۰۳/۰۲ تاریخ پذیرش: ۱۴۰۳/۰۷/۱۵

نوع مقاله: علمی‌ـ پژوهشی

چکیده. مدل‌های رگرسیون آمیخته به طور گسترده‌ای برای به دست آوردن روابط بین متغیر پاسخ و یک یا چند متغیر پیشبین از چندین گروه غیر همگن به کار می‌روند. از آنجایی بسیاری از داده‌ها در واقعیت دارای دم سنگینی هستند بنابراین استفاده از مدل‌های رگرسیون آمیخته معمولی با خطاهای نرمال می‌تواند باعث ایجاد انحراف در استنباط شود. کلاس توزیع‌های مخلوط مقیاسی نرمال بسیاری از توزیع‌های برجسته متقارن را به عنوان موارد خاص در بر می‌گیرد. در بسیاری از مطالعات از روش امید‌ـ ماکسیمم برای بدست آوردن تعداد خوشه‌ها و برآورد پارمترها استفاده می‌شود. در این مقاله یک رویکرد فازی برای بدست آوردن تعداد خوشه‌ها و برآورد پارامترها در مدل رگرسیون آمیخته با توزیع خطاهای مخلوط مقیاسی نرمال پیشنهاد شده است. دو مطالعه شبیه‌سازی برای بررسی ویژگی جانبی برآوردگرها و حساسیت مدل پیشنهادی در برابر نقاط پرت انجام شده است. همچنین یک مثال واقعی برای بررسی اثربخشی روش پیشنهادی انجام شده است.

2010 *Mathematics Subject Classification.* 62H30, 62J02, 90C70
E-mails: farzane.hashemi1367@kashanu.ac.ir.
عبارات و کلمات کلیدی. رگرسیون غیر خطی، کلاس‌بندی فازی، توزیع آمیخته متناهی، کلاس‌بندی ماکزیمم درستنمایی.

۱. مقدمه

مدل‌های رگرسیون آمیخته[1] (MRM) کاربردهای گسترده‌ای در بسیاری از زمینه‌ها از جمله مهندسی، زیست شناسی، بیومتریک، ژنتیک، پزشکی، اقتصاد سنجی، روانشناسی و بازاریابی دارند. این مدل‌ها برای بررسی رابطه بین متغیرهایی که از چندین گروه ناهمگن پنهان آمده‌اند استفاده می‌شوند. اگر متغیرهای پاسخ در مقابل متغیرهای پیشگو از یک تابع غیر خطی پیروی کنند از مدل‌های رگرسیون غیرخطی استفاده می‌شود. در حالت کلی یک مدل رگرسیون غیرخطی آمیخته با فرض توزیع خطاهای نرمال به صورت زیر مطرح می‌شود:

$$Y_j = \eta(\beta_i, x_i) + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma_i^2),\ j = 1, \ldots, n,\ i = 1, \ldots, g \tag{۱.۱}$$

با احتمال π_i، که در آن $Y = (Y_1, \ldots, Y_n)^\top$ متغیر پاسخ، η تابع رگرسیونی پیوسته و دوبار مشتق‌پذیر نسبت به پارامتر $\beta_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})^T$، π_iها وزن‌های آمیخته هستند به طوری که $\sum_{i=1}^{g} \pi_i = 1$ و x_j بردار متغیرهای پیشگوی متناظر با Y_j است. در مدل (۱.۱)، پارامترهای مدل را با $\Theta = \{\theta_1, \ldots, \theta_g, \pi\}$ که در آن $\pi = (\pi, \ldots, \pi_{g-1})$ و $\theta_i = (\beta_i, \sigma_i^2)$ تعریف می‌کنیم. Θ بردار پارامترهای مجهولی است که باید برآورد شوند. برآورد درستنمایی ماکزیمم Θ به طور کلی با مشخص کردن جواب تابع زیر به دست می‌آید:

$$\widehat{\Theta} = \arg\max_{\Theta} \sum_{j=1}^{n} \log \left(\sum_{i=1}^{g} \pi_i \phi \left(y_j; \eta(\beta_i, x_i), \sigma_i^2 \right) \right) \tag{۲.۱}$$

بطوریکه $\phi(.; \mu, \sigma^2)$ تابع چگالی نرمال با میانگین μ و واریانس σ^2 است. به هر حال تابع رابطه (۲.۱) غیر قابل حل است زیرا از مجموع چندین لگاریتم تشکیل شده است که این موضوع باعث بدست نیامدن روابط صریح برای پارامترهای مجهول می‌شود. بنابراین برای حل این مشکل با استفاده از الگوریتم امید-ماکسیمم[2] (EM) که توسط دمستر و همکاران [۷] معرفی شد، استفاده می کنیم. یاو و همکاران [۱۶] در چهارچوب داده کامل مدل رگرسیونی آمیخته کلاسیک را با استفاده از توزیع تی تعمیم دادند که اگر درجه آزادی در توزیع تی افزایش یابد به مدل نرمال نزدیک می‌شود. پونزو و مک نیکلاس [۱۴] توزیع آمیخته متناهی نرمال آلوده شده را برای خطاهای رگرسیونی در نظر گرفتند و با استفاده از الگوریتم EM برآورد پارامترهای مدل را بدست آوردند.

[1] Mixture Regression Models (MRMs)
[2] Expectation and Maximum (EM)

بدست آوردن برآورد پارامترها همزمان با طبقه‌بندی داده‌ها به داخل خوشه‌های همگن انجام می‌شود. بنابراین می‌توان از روش‌های خوشه‌بندی استفاده کرد. مدل‌های آمیخته برای خوشه‌بندی داده‌ها مدلی نیرومند و محبوب هستند، که با استفاده از الگوریتم EM به برآورد پارامترها و تشخیص خوشه‌بندی می‌پردازند که در منابع [۱۵، ۱۴] مورد مطالعه قرار گرفته‌اند. اشکال اصلی این روش عدم تشخیص تعداد گروها و حساسیت بالا به نقاط اولیه است که مانع استفاده از آن با مجموعه داده‌های بزرگ با تعداد زیادی خوشه ناهمگن می‌شود.

خوشه‌بندی فازی به طور گسترده‌ای در مسائل مختلف به کار می‌رود[۱۸]. به طور خاص، مرز بین دو بخش مجاور اغلب نامشخص است، و از این رو خوشه‌بندی فازی اجازه می‌دهد داده‌ها به بیش از یک خوشه مربوط شوند. در حالی که خوشه‌بندی فازی به طور گسترده در بسیاری از مسائل مورد توجه قرار گرفته است، کمتر برای مشخص کردن تعداد خوشه‌ها در مدل‌های رگرسیون آمیخته استفاده شده است. هاشمی و همکاران [۱۰] با استفاده از مجموعه‌های فازی برآورد پارامترها و تعداد خوشه‌های ناهمگن را در مدل آمیخته متناهی بیرنبام ساندرز[1] مورد ارزیابی قرار دادند. آنها همچنین نشان دادند سرعت همگرایی الگوریتم فازی در مقابل الگوریتم EM بیشتر است. در همین راستا هاشمی [۲] شرح مفصلی از خوشه‌بندی فازی را رساله دکتری آورده است. علاوه بر این، بیگدلی و دریجانی [۱] یک الگوریتم کلاس‌بندی فازی براساس توزیع آمیخته متناهی مخلوط مقیاسی نرمال چند متغیره ارائه دادند و خوشه‌بندی فازی را در این چارچوب تعبیه کردند. آنها دریافتند که روش درستنمایی فازی بهتر و کارآمدتر از سایر روش‌های عمل می‌کند. همچنین لو و چانگ [۱۱] از روش خوشه‌بندی فازی در رگرسیون تکه‌ای[2] استفاده کردند.

موارد بیان شده در قبل می‌تواند انگیزه‌ای برای معرفی یک الگوریتم طبقه‌بندی ماکسیمم درستنمایی فازی[3] (FCML) برای برآورد پارامترهای مدل رگرسیون آمیخته در جهت بدست آوردن تعداد کلاس‌ها و برآورد پارامترهای رگرسیون به طور همزمان باشد. درباره اهمیت استفاده از توزیع مخلوط مقیاسی نرمال[4] (SMN) می‌توان به خواصی از این توزیع در بعضی حالت‌های خاص آن برای شناسایی نقاط پرت نام برد. از دیگر موارد انگیزشی و نوآوری می‌توان به برآورد تعداد خوشه‌ها بدون در نظر گرفتن مقداری ثابت برای آنها اشاره کرد. با توجه به

[1] Finite Mixture of Birnbaum-Saunders
[2] Piecewise Regression
[3] Fuzzy Classification Maximum Likelihood
[4] Scale Mixture Normal

مزیت مدل‌های آمیخته برای خوشه‌بندی داده‌ها و برتری رویکرد فازی در تشخیص مرز بین خوشه‌ها، ما یک الگوریتم FCML برای مدل رگرسیون آمیخته ایجاد می‌کنیم.

ساختار مقاله به شرح زیر است. بخش ۲ به طور خلاصه توزیع SMN را بررسی و برخی از ویژگی‌های مرتبط با آن را تشریح می‌دهیم. بخش ۳ مدل پیشنهادی را بیان و یک روش مبتنی بر FCML را برای برآورد پارامترها و همچنین تخمین تعداد خوشه‌ها ارائه می‌کنیم. دو مطالعه شبیه‌سازی در بخش ۴ و یک مثال واقعی در بخش ۵ برای بررسی اثربخشی روش پیشنهادی انجام شده است.

۲. مروری بر مدل آمیخته رگرسیونی برپایه‌ی توزیع مخلوط مقیاسی نرمال

در این بخش ابتدا مروری بر چندین حالت خاص از توزیع SMN و سپس به تحلیل مدل آمیخته رگرسیونی برپایه‌ی این توزیع خواهیم پرداخت.

۱.۲. مروری بر توزیع مخلوط مقیاسی نرمال. در این بخش برخی نتایج اولیه و نمایش تصادفی از توزیع SMN را ارائه می‌دهیم.

متغیر تصادفی X دارای توزیع SMN با پارامتر مکان μ، پارامتر مقیاسی σ^2 و پارامتر ν که با نماد $\text{SMN}(\mu, \sigma^2, \nu)$ معرفی می‌شود، هرگاه نمایش تصادفی آن به صورت زیر باشد:

$$X \stackrel{d}{=} \mu + W^{-1/2} X_1, \quad X_1 \perp W, \tag{۱.۲}$$

که در آن $\perp$ علامت استقلال دو متغیر، X_1 یک متغیر تصادفی دارای توزیع نرمال با میانگین ۰ و واریانس σ^2 و W یک متغیر تصادفی مثبت مستقل از X_1 بوده که دارای تابع توزیع $H(.;\nu)$ است. با توجه به نمایش تصادفی (۱.۲) به سادگی می‌توان نتیجه گرفت که

$$X \Big| W = w \sim N(\mu, w^{-1}\sigma^2),$$

است. بنابراین می‌توان تابع چگالی متغیر تصادفی X را به صورت

$$f_{\text{SMN}}(x; \mu, \sigma^2, \nu) = \int_0^\infty \phi(x; \mu, w^{-1}\sigma^2) dH(w; \nu) \quad x \in \mathbb{R},$$

بدست آورد. با توزیع‌های مختلفی که برای W می‌توان در نظر گرفت موارد خاص از توزیع کلی SMN را بدست می‌آید. در این مقاله روی چند مورد رایج از موارد خاص توزیع SMN یعنی توزیع نرمال (N)، تی (t)، اسلش (SL) و نرمال آلوده شده (CN) متمرکز هستیم.

۲.۲. تحلیل مدل رگرسیون آمیخته با خطاهای مخلوط مقیاسی نرمال. در این بخش مدل رگرسیونی آمیخته با فرض پیروی خطاهای مدل از توزیع SMN مورد بررسی قرار می‌گیرد. اگر در رابطه (۱.۱) خطای مدل از توزیع SMN پیروی کند، آنگاه متغیر پاسخ دارای توزیع

$$Y_j \sim \text{SMN}(\eta(\beta_i, x_{ij}), \sigma_i^۲, \nu_i),\ j = ۱, \ldots, n,\ i = ۱, \ldots, g, \tag{۲.۲}$$

با احتمال π_i است. حال پارامترهای مدل را با $\Theta = \{\theta_۱, \ldots, \theta_g, \pi\}$ که در آن $\pi = (\pi, \ldots, \pi_{g-۱})$ و $\theta_i = (\beta_i, \sigma_i^۲, \nu_i)$ تعریف می‌کنیم. در اینصورت لگاریتم تابع درستنمایی y به صورت زیر خواهد بود:

$$\ell(\Theta|y) = \sum_{j=۱}^{n} \log \left(\sum_{i=۱}^{g} \pi_i \, f_{\text{SMN}}(y_j; \eta(\beta_i, x_j), \sigma_i^۲, \nu_i) \right).$$

ماکسیمم‌سازی مستقیم این نوع از توابع درستنمایی دشوار می‌باشد که یک روش جایگزین برای آن الگوریتم EM است. برای برآورد پارامترهای مدل معرفی شده، در این بخش ما از الگوریتم ECM [۱] استفاده می‌کنیم. بدین منظور ابتدا متغیر تخصیص[۲] را به صورت $Z_j = (Z_{۱j}, ..., Z_{gj})$ در نظر بگیرید که در آن برای $j = ۱, ..., n$ هر یک از Z_{ij} برابر ۱ خواهد بود اگر y_j متعلق به iامین ترکیب باشد و در غیر اینصورت $Z_{ij} = ۰$. در نتیجه می‌توان بیان کرد که Z_jها مستقل از یکدیگر دارای توزیع چند جمله‌ای[۳] با یک آزمایش و احتمالات $(\pi_۱, ..., \pi_g)$ خواهد بود که با $Z_j \sim \text{Multi}(۱; \pi_۱, ..., \pi_g)$ نشان داده می‌شود. بنابراین با استفاده از نمایش تصادفی (۱.۲) که در آن $W \sim H(w_j; \nu_i)$، می‌توان نمایش تصادفی سلسله مراتبی زیر را برای رابطه (۲.۲) در نظر گرفت:

$$\begin{aligned} Y_j | W_j = w_j, Z_{ij} = ۱ &\sim \text{N}\left(\eta(\beta_i, x_j), w_j^{-۱}\sigma^۲\right), \\ W_j | Z_{ij} = ۱ &\sim H(w_j; \nu_i), \\ Z_j &\sim \text{Multi}(۱; \pi_۱, \ldots, \pi_g). \end{aligned}$$

[1] Expectation Conditional Maximum
[2] Component
[3] Multinomial Distribution

در اینصورت لگاریتم تابع درستنمایی کامل برای Θ متناظر با مشاهدات y، بردار متغیر پنهان $w = (w_۱, ..., w_n)$ و بردار متغیر تخصیص $z_j = (z_{۱j}, ..., z_{gj})$ به صورت

$$\ell_c(\Theta|y, w, z) = \sum_{j=۱}^{n}\sum_{i=۱}^{g} z_{ij}\Bigg\{ \log \pi_i + \log h(w_j; \nu_i) - \frac{۱}{۲}\log \sigma_i^۲ - \frac{w_j}{۲\sigma_i^۲}\left(y_j - \eta(\beta_i, x_j)\right)^۲ \Bigg\},$$

خواهد بود، که در آن از مقادیر ثابت و مستقل از پارامترهای مجهول صرف نظر شده است. حال با محاسبه‌ی امید ریاضی لگاریتم تابع درستنمایی کامل نسبت به چگالی شرطی $(W, Z)|Y = y$ تابع محوری زیر به دست خواهد آمد:

$$Q(\Theta|\widehat{\Theta}^{(k)}) = \sum_{j=۱}^{n}\sum_{i=۱}^{g} \widehat{z}_{ij}^{(k)}\Bigg\{ \log \pi_i + \widehat{\Psi}_{ij}^{(k)} - \frac{۱}{۲}\log \sigma_i^۲ - \frac{\widehat{w}_{ij}^{(k)}}{۲\sigma_i^۲}\left(y_j - \eta(\beta_i, x_j)\right)^۲ \Bigg\}, \quad (۳.۲)$$

که در آن دو امید ریاضی شرطی

$$\hat{w}_{ij} = E[W_j|y_j, Z_{ij} = ۱, \widehat{\Theta}^{(k)}] \quad \text{و} \quad \hat{\Psi}_{ij} = E[h(w_j; \nu_i)|y_j, Z_{ij} = ۱, \widehat{\Theta}^{(k)}],$$

از روابط بخش ۱.۲ به دست می‌آیند.

در الگوریتم ECM برای برآورد ماکسیمم درستنمایی پارامترهای مدل آمیخته رگرسیون با خطاهای SMN، گام‌های E و CM زیر تکرار می‌شود:

- گام E: در هر تکرار، در این گام تابع محوری Q با در نظر گرفتن $\Theta = \widehat{\Theta}^{(k)}$ و محاسبه‌ی

$$\hat{z}_{ij}^{(k)} = \frac{\hat{\pi}_i^{(k)} f_{\text{SMN}}(y_j; \eta(\hat{\beta}_i^{(k)}, x_j), \hat{\sigma}_i^{(k)^۲}, \hat{\nu}_i^{(k)})}{\sum_{i=۱}^{g} \hat{\pi}_i^{(k)} f_{\text{SMN}}(y_j; \eta(\hat{\beta}_i^{(k)}, x_j), \hat{\sigma}_i^{(k)^۲}, \hat{\nu}_i^{(k)})},$$

و $\hat{w}_{ij}^{(k)}$ با استفاده از مشخص کردن حالت خاص توزیع SMN به دست می‌آید.

- گام CM: در این گام برای هر تکرار k، پارامتر مجهول Θ به صورت $\widehat{\Theta}^{(k+۱)}$ به روزرسانی می‌شود. برای به دست آوردن $\widehat{\Theta}^{(k+۱)}$ ابتدا فرض کنید که تابع

رگرسیونی $\eta(\beta_i, x_j)$ خطی باشد، یعنی فرض می‌کنیم $\eta(\beta_i, x_j) = x_j^\top \beta_i$. حال با ماکسیمم‌سازی (۳.۲) نسبت به Θ برآوردهای به روز شده به صورت

$$\begin{aligned}
\widehat{\beta}_i^{(k+1)} &= \left[\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{w}_{ij}^{(k)} x_j x_j^\top\right]^{-1} \left[\sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \hat{w}_{ij}^{(k)} y_j x_j\right], \\
\widehat{\sigma}^{(k+1)^2} &= \frac{1}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}} \sum_{j=1}^{n} \left\{ \hat{z}_{ij}^{(k)} \hat{w}_{ij}^{(k)} \big(y_i - \eta(\widehat{\beta}^{(k+1)}, x_j)\big)^2 \right\},
\end{aligned}$$

به دست می‌آیند. چون فرم بسته‌ای برای برآورد پارامتر ν_i وجود ندارد، آن را بوسیله ماکسیمم‌سازی تابع درستنمایی و به صورت زیر به روز رسانی می‌کنیم:

$$\widehat{\nu}_i^{(k+1)} = \arg\max_{\nu_i} \left\{ \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} f_{\text{SMN}}\Big(y_i; x_i^\top \widehat{\beta}^{(k+1)}, \widehat{\sigma}^{2(k+1)}, \nu_i\Big) \right\}.$$

برای تابع رگرسیونی دلخواه $\eta(\beta_i, x_j)$، ابتدا با استفاده از بسط تیلور مدل (۱.۱) حول $\widehat{\beta}_i^{(k)}$ داریم:

$$y_j = \eta\big(\widehat{\beta}_i^{(k)}, x_j\big) + \eta'\big(\widehat{\beta}_i^{(k)}, x_j\big)\big(\beta_i - \widehat{\beta}_i^{(k)}\big) + \epsilon_j.$$

که در آن $\eta'\big(\widehat{\beta}_j^{(k)}, x_j\big)$ مشتق اول تابع $\eta\big(\beta_i, x_j\big)$ نسبت به β_i و محاسبه شده در $\beta_i = \widehat{\beta}_i^{(k)}$ است. بنابراین مدل به صورت زیر تبدیل خواهد شد:

$$y_j - \eta\big(\widehat{\beta}_i^{(k)}, x_j\big) + \eta'\big(\widehat{\beta}_j^{(k)}, x_j\big)\widehat{\beta}_i^{(k)} = \eta'\big(\widehat{\beta}_i^{(k)}, x_j\big)\beta_j + \epsilon_j.$$

با تعریف متغیر پاسخ کاذب[1] به صورت $\widetilde{y}_j = y_j - \eta\big(\widehat{\beta}_i^{(k)}, x_j\big) + \eta'\big(\widehat{\beta}_i^{(k)}, x_j\big)\widehat{\beta}_i^{(k)}$ و متغیر پیش‌بین کاذب به صورت $\widetilde{x}_j = \eta'\big(\widehat{\beta}_j^{(k)}, x_j\big)$ مدل دوباره به یک مدل خطی تبدیل می‌شود. بنابراین پارامتر β_i با درنظر گرفتن متغیرهای کاذب و مدل رگرسیونی خطی همانند قبل برآورد خواهد شد.

این فرآیند تا زمانی تکرار می‌شود که یک شرط همگرایی برقرار شود. از رابطه $|\ell(\widehat{\Theta}^{(k+1)}) - \ell(\widehat{\Theta}^{(k)})| < ۱۰^{-۶}$ به عنوان یک شرط همگرایی استفاده می‌کنیم. حالت دیگر برای خاتمه فرآیند این است که تعداد تکرار به عدد ۵۰۰۰ برسد که در این صورت فرآیند متوقف می‌شود. همچنین انتخاب مقادیر اولیه را برای الگوریتم ECM می‌توان به صورت زیر خلاصه کرد:

[1] Pseudo

- ابتدا نمونه‌ی تحت مطالعه را با استفاده از روش K-means، به g گروه مجزا تقسیم می‌نماییم.
- با محاسبه‌ی نسبت‌های نقاط تعلق گرفته به هر گروه، مقادیر اولیه برای $\hat{\pi}_i^{(0)}$ را برابر نسبت iامین گروه در نظر می‌گیریم.
- برای هر i، مقادیر اولیه برای $\hat{\beta}_i^{(0)}$ و $\hat{\sigma}_i^{2(0)}$ را با استفاده از روش دستور nlm در نرم افزار R بدست می‌آوریم و مقدار $\hat{\nu}_i^{(0)}$ مطابق هر توزیع خاص در توزیع SMN در نظر گرفته می‌شود.

۳. کلاس‌بندی فازی برای مدل رگرسیون آمیخته با خطاهای مخلوط مقیاسی نرمال

الگوریتم FCML یک روش خوشه‌بندی داده‌ها است که ترکیبی از روش خوشه‌بندی کلاسیک فازی [۴] و مدل‌های مخلوط است. در همین راستا، برای یک نمونه تصادفی به اندازه n، از مدل رگرسیونی آمیخته با متغیر Y، می‌توان افراز $P = (P_1, \ldots, P_g)$ را برای مشاهدات $y = (y_1, \ldots, y_n)$ درنظر گرفت. این افراز می‌تواند به صورت $P_1 \cup \ldots \cup P_g = y$ نیز نمایش داده شود. حال می‌توان تابع نشانگر عضویت $u_1, \ldots, u_g$ را به صورت

$$u_k(y) = \begin{cases} 1 & \text{اگر} \quad y \in P_k; \\ & \\ 0 & \text{اگر} \quad y \notin P_k; \end{cases} \qquad \forall y \in y, \qquad k = 1, \ldots, g. \quad ,$$

بیان کرد.

مجموعه $u = (u_1, \ldots, u_g)$ در مشخص کردن عضویت همانند $z = (z_1, \ldots, z_g)$ در الگوریتم EM عمل می‌کند و این مجموعه را افراز $c-$سخت[1] از y می‌نامند. با استفاده از روش مکلاکلن و بسفورد [2] [۱۲] تابع لگاریتم درستنمایی برای مدل SMN-MRM با توجه به مجموعه $u = (u_1, \ldots, u_g)$ برای بدست آوردن کلاس‌بندی ماکسیمم درستنمایی

[1] c-hard
[2] Mclachlan and Basford

(CML) به صورت

$$\begin{aligned}
&\ell_{CML}(u,\Theta|y)\\
&=\sum_{i=۱}^{g}\sum_{y_j\in P_i}\Big\{\log\pi_i+\widehat{\Psi}_{ij}^{(k)}-\frac{۱}{۲}\log\sigma_i^{۲}-\frac{\widehat{w}_{ij}^{(k)}}{۲\sigma_i^{۲}}\left(y_j-\eta(\beta_i,x_j)\right)^{۲}\Big\}\\
&=\sum_{i=۱}^{g}\sum_{j=۱}^{n}u_i(y_j)\Big\{\log\pi_i+\widehat{\Psi}_{ij}^{(k)}-\frac{۱}{۲}\log\sigma_i^{۲}-\frac{\widehat{w}_{ij}^{(k)}}{۲\sigma_i^{۲}}\left(y_j-\eta(\beta_i,x_j)\right)^{۲}\Big\}. \qquad (۱.۳)
\end{aligned}$$

ماکسیمم کردن رابطه (۱.۳) منجر به بدست آوردن برآورد پارامترهای مجهول مدل می‌شود. همان طور که در الگوریتم EM بیان شد، برای $j=۱,...,n$ هر یک از z_{ij} برابر ۱ خواهد بود اگر y_j متعلق به iامین ترکیب باشد و در غیر این صورت $z_{ij}=۰$ هستند اما در مجموعه $u=(u_۱,\ldots,u_g)$ در نظر می‌گیریم که $u_i(\cdot)$ بر روی بازه $[۰,۱]$ تعریف شده باشد به‌طوری که در آن $\sum_{i=۱}^{g}\pi_i=۱$ و $\sum_{i=۱}^{g}u_i(y_j)=۱$ برقرار باشد.

بنابراین تابع لگاریتم درستنمایی برای مدل SMN-MRM در حالت فازی به صورت

$$\begin{aligned}
\ell_{FCML}(u,\Theta|y)&=\sum_{i=۱}^{g}\sum_{j=۱}^{n}u_i(y_j)\log\pi_i\\
&+\sum_{i=۱}^{g}\sum_{j=۱}^{n}u_i(y_j)\Big\{\widehat{\Psi}_{ij}^{(k)}-\frac{۱}{۲}\log\sigma_i^{۲}-\frac{\widehat{w}_{ij}^{(k)}}{۲\sigma_i^{۲}}\left(y_j-\eta(\beta_i,x_j)\right)^{۲}\Big\} \qquad (۲.۳)
\end{aligned}$$

نمایش داده می‌شود. با ماکسیمم کردن تابع بیان شده در رابطه (۲.۳) می‌توان پارامترهای مجهول مدل را با الگوریتم FCML بدست آورد.

برای به‌وجود آوردن عملکرد بهتر برای این الگوریتم، پیشنهاد می‌دهیم جملات جریمه به مانند زیر به رابطه (۲.۳) اضافه شوند و می‌توان این رابطه را به‌صورت زیر بازنویسی کرد:

$$\begin{aligned}
\ell_{FCML}(u,\Theta|y)=&n\sum_{i=۱}^{g}\pi_i\log\pi_i-\sum_{i=۱}^{g}\sum_{j=۱}^{n}u_i(y_j)\log u_i(y_j)+\sum_{i=۱}^{g}\sum_{j=۱}^{n}u_i(y_j)\log\pi_i\\
&+\sum_{i=۱}^{g}\sum_{j=۱}^{n}u_i(y_j)\Big\{\widehat{\Psi}_{ij}^{(k)}-\frac{۱}{۲}\log\sigma_i^{۲}-\frac{\widehat{w}_{ij}^{(k)}}{۲\sigma_i^{۲}}\left(y_j-\eta(\beta_i,x_j)\right)^{۲}\Big\}
\end{aligned}$$

با اضافه کردن $\sum_{i=1}^{g} \pi_i = 1$ و $\sum_{i=1}^{g} u_i(y_j)$ همراه با ضرایب لاگرانژ به تابع $\ell_{FCML}(u, \Theta|x)$، خواهیم داشت :

$$\tilde{\ell}_{FCML}(u, \Theta, \lambda) = \ell_{FCML}(u, \Theta|y) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{g} u_i(y_j) - 1 \right) + \lambda_{n+1} \left(\sum_{i=1}^{g} \pi_i - 1 \right), \quad (3.3)$$

به‌طوری که $\lambda = (\lambda_1, \ldots, \lambda_{n+1})$ ضرایب لاگرانژ هستند. بنابراین برآوردگرهای پارامترهای رگرسیون (Θ) را می‌توان با مشتق گرفتن از تابع $\tilde{\ell}_{FCML}$ نسبت به هر پارامتر و قرار دادن آنها برابر با ۰ به دست آورد. مراحل زیر را برای انجام الگوریتم فازی بیان شده در نظر می‌گیریم.

مرحله ۱: مقادیر به روز شده تابع عضویت فازی $u_i(y_j)$ به صورت زیر محاسبه خواهند شد:

$$\hat{u}_i(y_j) = \exp\left\{ \left(\widehat{d}_{ij}^{(k)} + \log \hat{\pi}_i^{(k)} \right) \right\} \Big/ \sum_{i=1}^{g} \exp\left\{ \left(\widehat{d}_{ij}^{(k)} + \log \hat{\pi}_i^{(k)} \right) \right\}, \quad (4.3)$$

بطوریکه

$$\widehat{d}_{ij}^{(k)} = \left\{ \widehat{\Psi}_{ij}^{(k)} - \frac{1}{2} \log \widehat{\sigma}_i^{(k)^2} - \frac{\widehat{w}_{ij}^{(k)}}{2\widehat{\sigma}_i^{(k)^2}} \left(y_j - \eta(\widehat{\beta}_i^{(k)}, x_j) \right)^2 \right\},$$

است.

مرحله ۲: برای بدست آوردن مقدار بروز شده π_i، تابع (۳.۳) را نسبت به π_i با وجود شرط $\sum_{i=1}^{g} \pi_i = 1$ ماکسیمم می‌کنیم و بنابراین داریم:

$$\hat{\pi}_i^{(new)} = \frac{1}{n} \sum_{j=1}^{n} \hat{z}_{ij} + \hat{\pi}_i^{(old)} \left(\log \hat{\pi}_i^{(old)} - \sum_{i=1}^{g} \hat{\pi}_i^{(old)} \log \hat{\pi}_i^{(old)} \right).$$

با توجه به رابطه بالا تعداد جدید کلاس‌ها به صورت زیر بدست می‌آید:

$$g^{(new)} = g^{(old)} - \#\{\hat{\pi}_i^{(old)} : \ \hat{\pi}_i^{(old)} < 1/n, \ i \in \{1, \ldots, g^{(old)}\}\},$$

که در آن $\#\{\}$ نماد تعداد عناصر صدق کننده در شرط موجود است. حال برای اینکه $\hat{\pi}_i^{(new)}$ و $\hat{u}_i(y_j)$ بدست آمده شرایط اصلی خود را بر اساس تعداد کلاسهای جدید دارا باشند، تغییر زیر را روی آنها اعمال می‌کنیم.

$$\hat{\pi}_i^{(new)} = \frac{\hat{\pi}_i^{(new)}}{\hat{\pi}_1^{(new)} + \cdots + \hat{\pi}_{g^{(new)}}^{(new)}}$$

و

$$\hat{u}_i(y_j) = \frac{\hat{u}_i(y_j)}{\hat{u}_1(y_j) + \cdots + \hat{u}_{g^{(new)}}(y_j)},$$

بطوریکه $\hat{u}_1(y_j) + \cdots + \hat{u}_{g^{(new)}}(y_j) = 1$ و $\hat{\pi}_1^{(new)} + \cdots + \hat{\pi}_{g^{(new)}}^{(new)} = 1$ هستند.

مرحله ۳: در این مرحله همانند قبل فرض می‌کنیم که تابع رگرسیونی $\eta(\beta_i, x_j)$ خطی باشد، یعنی فرض می‌کنیم $\eta(\beta_i, x_j) = x_j^\top \beta_i$. با ماکسیمم کردن تابع رابطه (۳.۳) نسبت به β_i، σ_i^2 و ν_i داریم:

$$\widehat{\beta}_i^{(k+1)} = \left[\sum_{j=1}^{n} \hat{u}_i(y_j)\hat{w}_{ij}^{(k)} x_j x_j^\top\right]^{-1} \left[\sum_{i=1}^{n} \hat{u}_i(y_j)\hat{w}_{ij}^{(k)} y_j x_j\right].$$

$$\widehat{\sigma}^{(k+1)^2} = \frac{1}{\sum_{j=1}^{n} \hat{u}_i(y_j)} \sum_{j=1}^{n} \left\{\hat{u}_i(y_j)\hat{w}_{ij}^{(k)} \big(y_i - \eta(\widehat{\beta}^{(k+1)}, x_j)\big)^2\right\}$$

$$\hat{\nu}_i^{(k+1)} = \arg\max_{\nu_i} \tilde{\ell}_{FCML}(u, \Theta, \lambda)\big|_{\beta_i = \hat{\beta}_i^{(k+1)},\, \sigma_i^2 = \hat{\sigma}_i^{(k+1)^2}}$$

برای حالتی که $\nu = \nu_1 = \ldots = \nu_g$ هستند پارامتر ν با ماکزیمم سازی تابع زیر انجام می‌شود.

$$\hat{\nu}^{(k+1)} = \arg\max_{\nu} \sum_{j=1}^{n} \log\left[\sum_{i=1}^{g} \hat{\pi}_i^{(k)} f_{\text{SMN}}\Big(y_i; x_i^\top \widehat{\beta}^{(k+1)}, \widehat{\sigma}^{2(k+1)}, \nu\Big)\right]$$

در پایان برای تابع رگرسیونی دلخواه $\eta(\beta_i, x_j)$، با در نظر گرفتن تابع عضویت فازی به جای متغیر تخصیص همانند پایان بخش ۲.۲ عمل می‌کنیم. مراحل ۱ الی ۳ را تا حصول شرط همگرایی ادامه می‌دهیم.

۱.۳. انتخاب پارامتر اولیه و انتخاب مدل مناسب در الگوریتم FCML. در این بخش فنون اجرای الگوریتم FCML بر روی مدل SMN-MRM را مورد بررسی قرار می‌دهیم. این فنون شامل مقادیر اولیه، شرایط همگرایی و معیار مناسب بودن مدل است.

۱) مقادیر اولیه:

مجموعه مقادیر اولیه را بصورت

$$g^{(0)} = n,\ \hat{\pi}_i^{(0)} = 1/n,\ \hat{\beta}_i^{(0)} = x_i,\ \hat{\sigma}_i^{2(0)} = 1,$$

در نظر می‌گیریم و توجه داشته باشید که مقدار $\hat{\nu}_i^{(0)}$ مطابق هر توزیع خاص در توزیع SMN در نظر گرفته می‌شود.

۲) شرایط همگرایی:
اگر $||u^{(l)} - u^{(l-1)}|| < \epsilon$ شد، حلقه متوقف می‌شود. در این بخش به ازای خطای $\epsilon = ۱۰^{-۶}$ استفاده شده است.

۳) معیار مناسبت مدل:
به منظور بررسی و مقایسه‌ی مدل‌های آماری مطرح شده در این مقاله، از دو معیار AIC[۱] و BIC[۲] که به صورت زیر محاسبه می‌شوند، استفاده خواهد شد:

$$\text{AIC} = ۲m - ۲\ell_{\max} \quad \text{و} \quad \text{BIC} = m\log n - ۲\ell_{\max},$$

که در آن m تعداد پارامترهای برآورد شده و $\ell_{\max}$ لگاریتم ماکسیمم درستنمایی است. هرچه برای یک مدل این دو معیار کوچک‌تر باشد مدل مفروض برازش بهتری را بر روی داده‌ها صورت خواهد داد.

برای به دست آوردن تعداد مناسب اجزای آمیخته g، باید الگویتم FCML را تا زمان محاسبه g مناسب ادامه داده و با استفاده از معیارهایی مناسب بودن مقدار g را مورد بررسی قرار داد.

۴. مطالعات شبیه‌سازی

۱.۴. خواص مجانبی برآوردگرها. اولین مطالعه شبیه‌سازی برای بررسی عملکرد برآوردگرهای مبتنی بر روش فازی از طریق روش مبتنی بر روش‌های شرح داده شده در بخش ۳ انجام شده است. در این شبیه‌سازی ۵۰۰ نمونه ۱۰۰، ۵۰۰ و ۱۰۰۰تایی از مدل t-MRM با $g = ۲$ تحت مدل

$$y = \frac{\beta_۰}{۱ + \beta_۱ \exp\{-\beta_۲ x_۱ - \beta_۳ x_{۲j}\}} + \varepsilon$$

تولید می‌کنیم. پارامترهای مدل به صورت

$$\pi_۱ = ۰/۴,\ \beta_۱ = (۳۷, ۴۶, ۰/۷۵, ۲)^\top,\ \beta_۲ = (۴۲, ۵۱, ۱/۵, ۳)^\top,$$
$$\sigma_۱^۲ = ۱,\ \sigma_۲^۲ = ۲,\ \nu_۱ = ۵,\ \nu_۲ = ۱۰,$$

[1] Akaike Information Criterion
[2] Bayesian Information Criterion

جدول ۱: نتایج شبیه‌سازی برای خواص مجانبی برآوردگرهای توزیع t-MRM با استفاده از الگوریتم FCML و EM.

n	روش	اندازه	π_1	$\sigma_1^2(1)$	$\sigma_2^2(2)$	$\beta_{10}(37)$	$\beta_{11}(46)$	$\beta_{12}(0/75)$	$\beta_{13}(2)$	$\beta_{20}(42)$	$\beta_{21}(51)$	$\beta_{22}(1/5)$	$\beta_{23}(3)$	$\nu_1(5)$	$\nu_2(10)$
۱۰۰	FCML	MSE	۰/۰۹۵۸	۰/۰۵۲۲	۰/۰۷۰۱	۰/۰۲۷۳	۰/۰۴۳۰	۰/۰۴۴۵	۰/۰۵۳۴	۰/۰۱۱۸	۰/۰۲۲۱	۰/۰۵۶۳	۰/۰۶۱۲	۰/۵۶۳۰	۰/۷۷۱۰
		RB	۰/۰۲۰۹	۰/۰۲۴۶	۰/۰۲۸۴	۰/۰۹۳۰	۰/۱۰۳۰	۰/۱۳۶۴	۰/۱۴۶۷	۰/۱۱۳۶	۰/۱۴۲۳	۰/۱۴۹۲	۰/۱۵۳۹	۰/۸۸۹۲	۰/۹۹۲۶
	EM	MSE	۰/۱۰۲۴	۰/۰۶۲۹	۰/۰۸۹۵	۰/۰۳۸۲	۰/۰۵۱۱	۰/۰۴۹۰	۰/۰۵۳۷	۰/۰۱۲۳	۰/۰۱۶۹	۰/۰۶۰۱	۰/۰۶۳۰	۰/۵۸۷۴	۰/۷۸۶۴
		RB	۰/۰۲۴۰	۰/۰۲۵۴	۰/۰۳۰۸	۰/۱۰۴۷	۰/۱۱۹۸	۰/۱۴۹۳	۰/۱۵۴۲	۰/۱۳۸۱	۰/۱۵۹۲	۰/۱۵۲۶	۰/۱۶۰۱	۰/۹۱۲۰	۰/۹۹۸۹
۵۰۰	FCML	MSE	۰/۰۸۸۲	۰/۰۴۱۳	۰/۰۵۷۰	۰/۰۲۰۹	۰/۰۳۵۳	۰/۰۳۹۸	۰/۰۵۲۳	۰/۰۱۱۳	۰/۰۲۱۵	۰/۰۵۲۷	۰/۰۵۹۵	۰/۴۹۶۳	۰/۷۲۱۲
		RB	۰/۰۱۸۳	۰/۰۲۲۳	۰/۰۲۱۵	۰/۰۷۶۲	۰/۰۸۰۴	۰/۱۲۰۸	۰/۱۱۲۰	۰/۰۸۰۸	۰/۱۲۰۸	۰/۱۴۰۸	۰/۱۴۹۷	۰/۸۰۴۲	۰/۹۰۶۸
	EM	MSE	۰/۰۹۱۲	۰/۰۵۲۴	۰/۰۶۵۹	۰/۰۳۰۳	۰/۰۴۳۶	۰/۰۳۱۵	۰/۰۵۲۹	۰/۰۱۱۸	۰/۰۱۵۵	۰/۰۵۵۹	۰/۰۶۱۲	۰/۵۲۴۹	۰/۷۵۰۱
		RB	۰/۰۲۰۹	۰/۰۲۳۴	۰/۰۲۸۱	۰/۰۸۳۲	۰/۰۹۴۴	۰/۱۳۶۲	۰/۱۲۸۸	۰/۱۰۷۸	۰/۱۳۵۶	۰/۱۴۸۸	۰/۱۵۵۲	۰/۸۵۹۳	۰/۹۱۰۸
۱۰۰۰	FCML	MSE	۰/۰۶۹۵	۰/۰۳۱۰	۰/۰۴۲۱	۰/۰۱۸۹	۰/۰۱۹۸	۰/۰۳۷۳	۰/۰۵۱۵	۰/۰۱۰۷	۰/۰۲۰۶	۰/۰۵۱۸	۰/۰۵۴۸	۰/۴۴۰۲	۰/۶۸۵۷
		RB	۰/۰۱۵۴	۰/۰۲۱۲	۰/۰۱۹۲	۰/۰۵۴۰	۰/۰۵۷۹	۰/۱۰۰۶	۰/۱۰۰۶	۰/۰۷۶۱	۰/۱۱۵۲	۰/۱۳۵۶	۰/۱۴۵۷	۰/۷۵۶۹	۰/۸۵۴۹
	EM	MSE	۰/۰۷۳۷	۰/۰۴۵۳	۰/۰۵۷۰	۰/۰۲۶۴	۰/۰۲۷۲	۰/۰۳۹۱	۰/۰۵۲۰	۰/۰۱۱۱	۰/۰۱۴۹	۰/۰۵۴۰	۰/۰۵۷۸	۰/۴۷۴۹	۰/۶۹۹۷
		RB	۰/۰۱۹۶	۰/۰۲۲۱	۰/۰۲۵۴	۰/۰۶۴۹	۰/۰۷۲۸	۰/۱۱۲۶	۰/۱۱۷۳	۰/۰۹۶۶	۰/۱۲۷۶	۰/۱۳۹۱	۰/۱۵۰۶	۰/۸۰۴۲	۰/۸۸۴۰

هستند. همچنین متغیر پیشگوی $x_j = (x_{1j}, x_{2j})^\top$ نیز برای هر n یعنی x_{1j} و x_{2j} به ترتیب از توزیع یکنواخت با فاصله $U(0, 15)$ و $U(2, 10)$ تولید شده و ثابت در نظر گرفته شده است. در هر تکرار پس از برآورد پارامترها اریبی و مربع خطای آنها محاسبه شده است. میانگین اریبی نسبی و ریشه میانگین مربعات خطای به صورت زیر محاسبه می‌شوند:

$$\text{RB} = \frac{1}{500}\sum_{i=1}^{500}\left|\frac{\hat{\theta}_i - \theta_{true}}{\theta_{true}}\right| \quad \text{و} \quad \text{RMSE} = \sqrt{\frac{1}{500}\sum_{i=1}^{500}\left(\hat{\theta}_i - \theta_{true}\right)^2},$$

که در آن $\hat{\theta}_i$ برآورد پارامتر در تکرار iام و θ_{true} مقدار واقعی آن پارامتر است. در جدول ۱ نتایج این شبیه‌سازی خلاصه شده است. همانطور که از این جدول ملاحظه می‌شود، با افزایش حجم نمونه مقدار RB و MSE کاهش پیدا می‌کند. همچنین ملاحظه می‌شود که مقادیر RB و MSE برای پارامترهای مجهول در اکثر موارد الگوریتم FCML بهتر از الگوریتم EM است.

۲.۴. بررسی حساسیت برآوردگرهای فازی در حضور نقاط پرت. هدف این شبیه‌سازی بررسی کارایی مدل پیشنهادی ما برای خوشه‌بندی داده‌ها در حضور نقاط پرت است. در این شبیه‌سازی ۲۰۰ تکرار با اندازه نمونه $n = 500$ از مدل

$$y = \frac{1}{1 + \beta_1 \exp\{-\beta_2 x_1\}} + \varepsilon$$

با $g = ۳$ شبیه سازی می‌کنیم. فرض می‌شود خطاهای مدل از توزیع چوله تی[1]، به عنوان تعمیمی از توزیع چوله نرمال[2] که توسط آزالینی و واله [۳] معرفی شد، با دو نوع پارامتر درجه آزادی: سناریو اول (S۱) $\nu = (۵, ۱۰, ۱۲)$ و سناریو دوم (S۲) $\nu = (۱۴, ۱۸, ۲۱)$ باشند. همچنین متغیر پیشگوی $x = (۱, x_{j۱})^\top$ نیز برای هر n یعنی $x_{j۱}$ از فاصله‌ی $(۱, ۱۵)$ بطور یکنواخت تولید شده و ثابت در نظر گرفته شده است. پارامترهای مدل به صورت

$$\pi_۱ = \pi_۲ = \pi_۳ = ۱/۳,\ \beta_۱ = (-۳, ۱)^\top,\ \beta_۲ = (۳, ۱)^\top,\ \beta_۳ = (۰/۵, ۱)^\top,$$
$$\sigma_۱^۲ = ۰/۲,\ \sigma_۲^۲ = ۰/۴,\ \sigma_۳^۲ = ۰/۶,\ \lambda_۱ = ۱,\ \lambda_۲ = ۲,\ \lambda_۳ = ۱/۵,$$

هستند. به منظور بررسی تأثیر مشاهدات پرت در انتخاب مدل و عملکرد خوشه‌بندی، ۵۰ داده تولید شده از توزیع یکنواخت $U(-۱, ۴)$ را اضافه می‌کنیم و آنها را به صورت تصادفی به یکی از سه کلاس اختصاص می‌دهیم. در هر یک از ۲۰۰ نمونه تولید شده چهار حالت خاص SMN بر این داده ها با دو روش EM و FCML برازش داده می‌شوند. همچنین برای هر مدل و در هر الگوریتم مورد استفاده، دو شاخص اطلاعات متقابل تنظیم شده[3] (AMI) و شاخص رتبه‌ی اصلاح شده[4] (ARI) را مورد محاسبه قرار داده‌ایم. علاوه بر این، میانگین زمان CPU (بر حسب ثانیه) برای هر سناریو ثبت شده است. همانطور که در جدول ۲ نشان داده شده است، زمانی که عبارات خطا تحت سناریوی $S۱$ تولید می‌شوند، t-MRM عملکرد مناسب‌تری تحت روش برآوردیابی فازی نسبت به سایر همتایان ارائه می‌دهند. با افزودن ۵۰ نقطه پرت، t-MRM بهترین عملکرد مناسب و خوشه‌بندی را برای داده‌های تولید شده ارائه می‌دهد. برعکس، زمانی که خطاها تحت سناریوی $S۲$ (دم سنگین) تولید می‌شوند، CN-MRM برای مجموعه داده‌های غیر پرت و پرت از مدل‌های دیگر بهتر عمل می‌کنند. در مورد زمان CPU نیز همانطور که در جدول آورده شده است روش فازی به دلیل اینکه بطور خودکار تعداد خوشه‌ها را تشخیص می‌دهد بهتر از روش الگوریتم EM عمل می‌کند.

۵. داده‌های ایندومتاسین

در این بخش یک تجزیه و تحلیل داده ایندومتاسین برای بررسی کارایی مدل بیان شده ارائه می‌دهیم، که اغلب به عنوان یک نمونه در مدل سازی رگرسیون غیرخطی استفاده می‌شود.

[1] Skew-t (ST)
[2] Skew Normal
[3] Adjusted Mutual Information
[4] Adjusted Rank Index

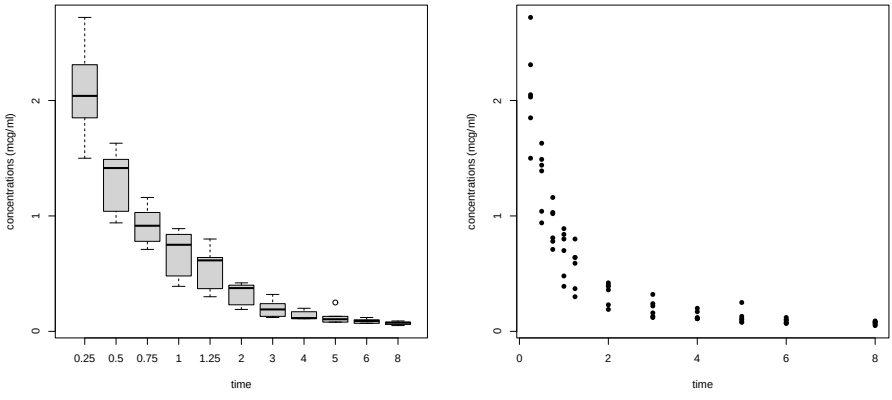
جدول ۲: نتایج شبیه‌سازی برای ارزیابی انتخاب مدل و عملکرد خوشه‌بندی چهار مدل SMN-MRM زمانی که داده‌ها با فرض توزیع خطاهای چوله تی بر اساس شرایط با و بدون افزودن نقاط نویز تحت الگوریتم‌های FCML و EM.

نویز	اندازه	CN-MRM		t-MRM		SL-MRM		N-MRM	
		EM	FCML	EM	FCML	EM	FCML	EM	FCML
۰	AIC	۲۰۴۳/۸۰۷	۱۹۵۱/۴۱۳	۲۰۴۲/۷۱۴	۱۹۵۰/۱۴۴	۲۰۴۹/۱۵۰	۱۹۵۶/۸۷۳	۲۰۷۱/۲۲۲	۱۹۸۳/۱۰۶
	BIC	۲۰۶۰/۲۲۶	۱۹۶۷/۸۳۲	۲۰۵۹/۱۳۳	۱۹۶۶/۵۶۳	۲۰۶۸/۲۰۷	۱۹۷۵/۹۳۱	۲۰۸۴/۹۲۴	۱۹۹۶/۸۰۸
	ARI	۰/۸۷۹	۰/۹۹۸	۰/۸۹۸	۰/۹۹۹	۰/۸۷۸	۰/۹۹۹	۰/۸۴۲	۰/۹۹۸
	AMI	۰/۷۶۴	۰/۹۹۴	۰/۷۶۳	۰/۹۹۴	۰/۷۶۲	۰/۹۹۴	۰/۷۵۵	۰/۹۹۴
	CPU Time	۱۲/۰۶۷	۳/۳۰۵	۱۰/۷۶۲	۳/۲۳۳	۱۵/۸۰۹	۵/۸۶۰	۱۶/۴۳۶	۷/۶۷۰
۵۰	AIC	۲۰۵۷/۶۶۳	۱۹۹۲/۶۷۸	۲۰۵۵/۸۴۲	۱۹۸۷/۶۶۸	۲۰۶۷/۹۰۳	۲۰۵۵/۶۵۴	۲۰۶۸/۵۲۱	۲۰۳۳/۷۶۴
	BIC	۲۰۷۵/۸۱۶	۲۰۱۰/۸۳۱	۲۰۷۲/۹۹۵	۲۰۰۵/۸۲۱	۲۰۸۸/۰۲۵	۲۰۷۷/۷۷۶	۲۰۸۳/۶۳۳	۲۰۴۸/۸۷۶
	ARI	۰/۷۰۲	۰/۹۲۶	۰/۷۰۵	۰/۹۲۹	۰/۷۰۰	۰/۹۰۵	۰/۶۹۹	۰/۹۲۱
	AMI	۰/۶۲۳	۰/۸۶۰	۰/۶۲۹	۰/۸۶۵	۰/۶۱۵	۰/۸۳۹	۰/۶۱۰	۰/۸۵۶
	CPU Time	۱۲/۲۰۶	۵/۰۴۵	۱۱/۷۵۲	۴/۳۰۳	۱۶/۹۱۲	۶/۸۲۵	۱۷/۵۰۷	۸/۸۳۷
۰	AIC	۱۹۷۵/۷۲۳	۱۹۵۶/۴۱۸	۱۹۷۷/۵۶۰	۱۹۵۹/۹۳۰	۲۰۰۴/۳۴۳	۱۹۸۵/۱۰۹	۱۹۹۳/۵۰۸	۱۹۶۷/۴۸۳
	BIC	۱۹۹۱/۱۴۳	۱۹۷۳/۸۳۸	۱۹۹۴/۹۸۰	۲۰۲۵/۳۴۹	۲۰۲۳/۴۰۰	۲۰۰۴/۱۶۶	۲۰۰۷/۲۱۰	۱۹۸۱/۱۸۵
	ARI	۰/۷۸۸	۰/۹۸۳	۰/۷۷۴	۰/۹۸۰	۰/۷۵۴	۰/۹۶۵	۰/۷۶۸	۰/۹۷۶
	AMI	۰/۷۰۱	۰/۹۶۹	۰/۶۹۷	۰/۹۶۶	۰/۶۷۴	۰/۹۴۱	۰/۶۹۱	۰/۹۶۱
	CPU Time	۱۳/۰۶۱	۵/۴۵۳	۱۴/۹۴۲	۶/۱۶۱	۱۶/۷۴۸	۷/۳۶۴	۱۹/۱۸۲۱	۸/۱۰۵
۵۰	AIC	۲۶۹۱/۰۲۰	۲۶۰۲/۵۴۷	۲۷۱۶/۹۲۱	۲۶۱۱/۶۷۳	۲۸۰۱/۶۲۶	۲۷۴۵/۱۵۱	۲۷۷۱/۰۴۴	۲۷۲۴/۲۲۴
	BIC	۲۷۱۱/۷۴۵	۲۷۰۱/۷۰۰	۲۷۳۴/۶۴۷	۲۷۲۰/۸۲۶	۲۸۲۴/۶۲۰	۲۷۸۱/۲۷۳	۲۷۸۷/۴۳۷	۲۷۳۹/۳۳۵
	ARI	۰/۶۲۵	۰/۶۵۰	۰/۶۲۲	۰/۶۴۷	۰/۶۱۴	۰/۶۴۸	۰/۶۲۰	۰/۶۴۰
	AMI	۰/۵۴۹	۰/۵۷۹	۰/۵۴۴	۰/۵۷۵	۰/۵۳۳	۰/۵۷۶	۰/۵۴۲	۰/۵۷۰
	CPU Time	۱۸/۷۹۸	۹/۹۲۷	۱۹/۰۶۴	۱۰/۴۹۶	۱۸/۶۹۲	۹/۲۶۳	۲۱/۹۴۵	۱۲/۸۲۶

این داده‌ها توسط افرادی مانند پینهیرو باتیس [۱۳] و داویدین [۶] مورد تجزیه و تحلیل قرار گرفتند. همچنین این مجموعه داده توسط گراسی [۸] مورد بررسی کلاس‌بندی قرار گرفتند. شکل ۱ نمودار پراکندگی و نمودارهای جعبه‌ای را برای غلظت پلاسمایی ایندومتاسین به زمان در داده‌ها نشان می‌دهد. هدف از این آزمایش این بود که ببینیم زمان (x) چگونه بر غلظت پلاسمایی ایندومتاسین (y) تأثیر می‌گذارد. با توجه به شکل ۱ می‌توان مشاهده کرد که چند الگوی غیرخطی ظاهری بین دو متغیر وجود دارد. علاوه بر این، نمودارهای جعبه‌ای وجود ویژگی‌های غیر نرمال و دم سنگینی را در داده‌ها نشان می‌دهد. برای تجزیه و تحلیل داده‌های ایندومتاسین، از مدل نمایی زیر استفاده کردیم:

$$y = \beta_0 \exp\{-\exp\{\beta_1 t\}\} + \varepsilon.$$

در ابتدا برای مشخص کردن g بهینه با استفاده از الگوریتم EM برای مقادیر متفاوت از g این الگوریتم را پیاده‌سازی می‌کنیم و با استفاده از مقدار BIC بهترین تعداد کلاس را انتخاب



شکل ۱: نمودارهای جعبه‌ای از غلظت ایندومتاسین بر اساس موقعیت اندازه‌گیری (سمت چپ) و نمودار پراکندگی غلظت دارو مشروط به زمان از زمان تزریق (راست) با استفاده از داده‌های ایندومتاسین.

جدول ۳: مقایسه مدل برای برازش MRM های مختلف به داده‌های ایندومتاسین بر اساس معیار BIC.

مدل	$g=۱$	$g=۲$	$g=۳$	$g=۴$	$g=۵$	$g=۶$
N-MRM	۸۸۳/۸۴۶	۸۷۷/۳۰۴	۸۶۹/۱۹۷	۸۶۱/۲۶۱	۸۷۲/۱۷۱	۹۳۱/۳۸۸
SL-MRM	۷۷۹/۶۴۰	۷۶۹/۰۹۹	۷۵۵/۹۹۲	۷۶۱/۰۵۶	۷۷۳/۳۹۲	۸۳۲/۱۴۹
t-MRM	۷۴۳/۸۶۲	۷۳۳/۳۲۰	۷۲۵/۲۱۴	۷۲۷/۲۷۷	۷۳۹/۵۳۷	۷۷۴/۱۰۵
CN-MRM	۷۶۳/۹۳۰	۷۵۶/۳۸۸	۷۳۷/۲۸۱	۷۴۸/۳۴۵	۷۵۹/۱۲۲	۸۰۱/۲۰۷

می‌کنیم. جدول ۳ نتایج برازش در الگوریتم EM برای چهار مدل کاندید در کلاس SMN-MRM برای مقادیر مختلف g نمایش داده است. با توجه به جدول ۳ مشاهده شده است که در تمامی مدل‌ها به جز مدل N-MRM مقدار $g=۳$ بهینه‌ترین تعداد کلاس بر اساس معیار BIC است. بنابراین در استفاده از الگوریتم EM بهترین تعداد کلاس برابر با ۳ است که در مدل t-MRM مشاهده شده است. یکی از تفاوت‌های اساسی در استفاده از الگوریتم FCML در مقابل الگوریتم EM، شناسایی تعداد کلاس بهینه در اجرای یک بار الگوریتم است در صورتی که در الگوریتم EM مشاهده شد این مقدار بهینه با استفاده از مقایسه چند مقدار مختلف حاصل می‌شود. این موضوع باعث افزاش سرعت اجرا و بدست آوردن بهترین

جدول ۴: مقایسه مقادیر BIC برای برازش MRM های در دو الگوریتم FCML و EM. مختلف به داده‌های ایندومتاسین.

الگوریتم		N-MRM	SL-MRM	t-MRM	CN-MRM
EM	BIC	۸۶۱/۲۶۱	۷۵۵/۹۹۲	۷۲۵/۲۱۴	۷۳۷/۲۸۱
	time CPU	۹/۴۲	۱۲/۳۳	۱۰/۱۹	۹/۱۵
	g	۴	۳	۳	۳
FCML	BIC	۸۵۲/۱۳۷	۷۴۶/۹۸۲	۷۱۸/۱۶۵	۷۳۷/۴۲۰
	time CPU	۴/۹۳	۴/۸۵	۵/۹۶	۵/۶۰
	g	۳	۳	۳	۳

g ممکن در کمترین زمان است. در ادامه نتایج مربوط به مقایسه دو الگوریتم برای چهار مدل مختلف SMN-MRM در جدول ۴ آورده شده است. نتایج نشان می‌دهد که سه مدل t-MRM، SL-MRM و CN-MRM همگی برتر از N-MRM هستند، که دلیل آن به خاطر عدم وجود دم سنگینی در مدل مذکور است. t-MRM بهترین کلاس‌بندی را در بین هر دو الگوریتم برای داده‌های ایندومتاسین فراهم کرده است. همچنین تفاوت زمان اجرا بین الگوریتم FCML و EM نشان می‌دهد این استفاده از الگوریتم فازی بهتر است زیرا هم زمان اجرا کمتری نیاز دارد و در اکثر موارد مقادیر BIC کمتری دارد.

برای بررسی تأثیر مشاهدات پرت بر تخمین پارامترهای مدل، فاصله درستنمایی[1] (LD) را محاسبه می‌کنیم [۵]، که به صورت زیر تعریف شده است:

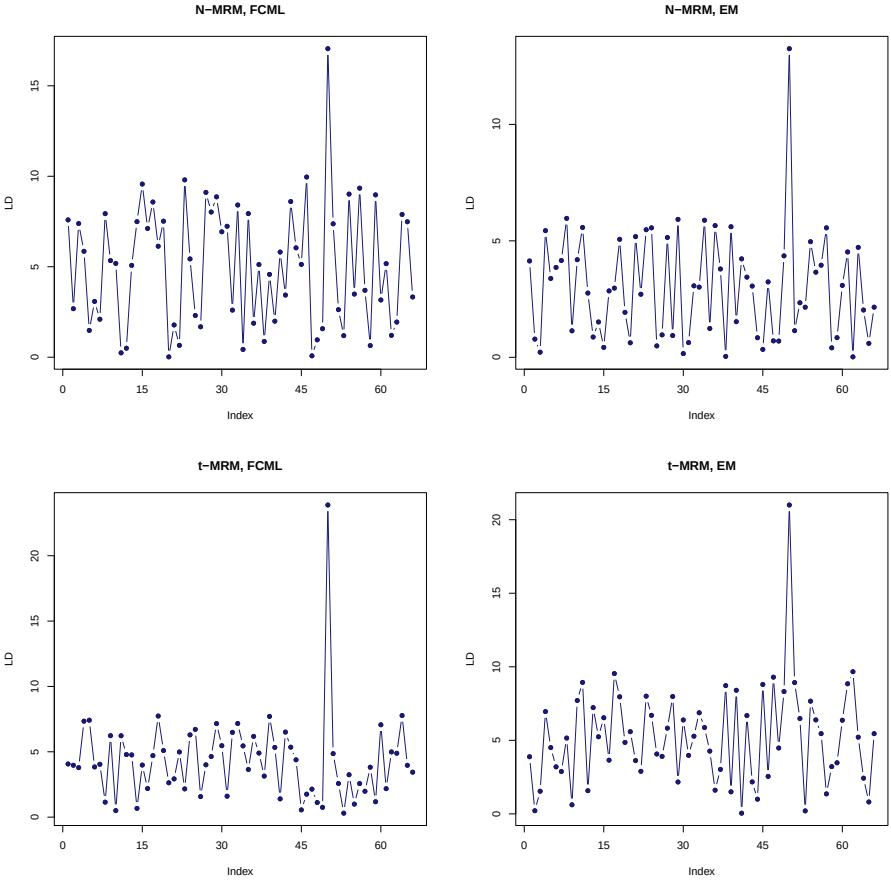
$$LD_s = ۲|\ell(\hat{\Theta}|y) - \ell(\hat{\Theta}_{-s}|y)|, \qquad (۱.۵)$$

که در آن $\hat{\Theta}_{-s}$ برآورد پارامترهای مدل بعد از حذف مشاهده sام می‌باشد. اگر مقدار LD نسبتاً زیاد باشد، آن مشاهده را به عنوان یک داده پرت تعیین در نظر می‌گیریم.

شکل ۲ مقادیر LD را نشان می‌دهد که با استفاده از (۱.۵) تحت دو مدل N-MRM و t-MRM برای دو الگوریتم EM و FCML بدست آمده‌اند. همان طور که در شکل ۲ مشاهده می‌شود، مدل t-MRM با استفاده از دو الگوریتم مورد #۵۰ را به عنوان یک نقطه پرت مشخص می‌کنند، در صورتی که این مشاهده در الگوریتم EM برای مدل N-MRM یک نقطه پرت ملایم در نظر گرفته شده است. شواهد تجربی نشان می‌دهد که مدل t-MRM

[1] likelihood Distance

بهترین برازش می‌تواند در برابر مشاهدات بسیار تأثیرگذار که همان نقاط پرت هستند داشته باشد.



شکل ۲: نمودار مقادیر LD برای داده‌های ایندومتاسین که توسط دو مدل N-MRM و t-MRM برای دو الگوریتم EM و FCML.

۶. نتیجه‌گیری

در این مقاله ضمن بیان مدل رگرسیون آمیخته برای توزیع SMN و بررسی برخی ویژگی‌های آن‌ها، برآورد پارامتر و کلاس‌بندی داده‌ها بر اساس الگوریتم FCML بیان و ویژگی‌های آنها مورد تجزیه و تحلیل قرار گرفت. همانطور که با مثال واقعی و شبیه‌سازی نشان داده شده است،

استفاده از الگوریتم FCML می‌تواند به عنوان یک رقیب و جایگزین برای سایر الگوریتم‌های پیشنهادی برای کلاس‌بندی مطرح شده در سال‌های اخیر باشد. همچنین مطالعات شبیه‌سازی انجام شده تحت این الگوریتم در مقابل الگوریتم EM نشان می‌دهد که کلاس‌بندی با این شیوه دارای ویژگی‌های مطلوب در برازش کلاس صحیح داده‌ها می‌باشد.

با توجه به نتایج بدست آمده از مزایای استفاده از الگوریتم FCML بدست آوردن تعداد خوشه‌ها است که منجر به زمان کمتر برای انجام الگوریتم می‌شود. باید توجه داشت که استفاده از این الگوریتم در بعضی موارد می‌تواند از لحاظ معیار BIC مناسب نباشد و الگوریتم EM مناسب‌تر باشد. با این حال استفاده از الگوریتم EM در حالتی که تعداد کلاس صحیح داده‌ها مشخص نباشد مناسب نیست زیرا در این الگوریم نیاز است با شاخص‌هایی ابتدا تعداد کلاس صحیح را بدست آورد که این موضوع نیاز به وقت زیاد دارد.

برای آینده تحقیق می‌توان این الگوریتم را برروی آمیخته متناهی از اثر مختلط خطی[1] پیاده‌سازی کرد. از موارد دیگر که به آن می‌توان اشاره کرد تعمیم الگوریتم پیشنهادی در حالتی که خطاهای مدل از توزیع چوله نرمال معرفی شده توسط آزالینی و واله [۳] پیروی کنند، است.

مراجع

[۱] بیگدلی، ح، و دریجانی، س. (۱۴۰۱) یادگیری مبتنی بر کلاس‌بندی $c-$ میانگین با استفاده از توزیع آمیخته مقیاسی نرمال با اطلاعات گمشده. سیستم های فازی و کاربردها، شماره ۵، صص. ۲۰۳ تا ۲۲۵.

[۲] هاشمی، ف. (۱۳۹۸) تحلیل عاملی کلاسیک و فازی با استفاده از توزیع‌های چوله متقارن و آمیخته آن‌ها. رساله دکتری، دانشگاه شهید باهنر.

[3] A. Azzalini and A. D. Valle. The multivariate skew-normal distribution. Biome., vol. 83 (4), pp. 715–726.

[4] J.C. Bezdek. “Pattern Recognition with Fuzzy Objective Function Algorithms”. Klu. Aca. Publ., Norwell, MA, USA, 1981.

[5] R.D. Cook and S. Weisberg. “Residuals and Influence in Regression”. Chap and Hall, New York, 1982.

[6] M. Davidian. “Nonlinear models for repeated measurement data”. Routledge, 2017.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. J. Roy. Stat. Soc.: Ser. B, vol. 39, 1977. pp. 1–22.

[8] M. Geraci. “Modelling and estimation of nonlinear quantile regression with clustered data”. Comp. Stat. Dat. Anal., vol. 136, 2019, pp. 30–46.

[1] Linear Mixed Effect

[9] D. E. Gustafson and W.C. Kessel. “Fuzzy clustering with a fuzzy covariance matrix”. IEEE conf. on deci. and conto. inclu. the 17th sympo. on adap. proc., 1979, pp. 761–766.

[10] F. Hashemi, M. Naderi and M. Mashinchi. “Clustering right-skewed data stream via Birnbaum–Saunders mixture models: A flexible approach based on fuzzy clustering algorithm”. App. Sof. Comp., vol. 82, 2019, 105539.

[11] K.P. Lu and S.T. Chang. “A fuzzy classification approach to piecewise regression models”. App. Sof. Comp., vol. 69, 2018, pp. 671–688.

[12] G. J. McLachlan and K.E.Basford. “Mixture Models: Inference and Applications to Clustering”. vol. 84, Marcel Dekker, 1988.

[13] J.C. Pinheiro and D.M. Bates. “Mixed-Effects Models in S and S-PLUS”. Springer Verlag, New York, 2000.

[14] A. Punzo and P.D. McNicholas. “Robust Clustering in Regression Analysis via the Contaminated Gaussian Cluster-Weighted Model”. J. Class., vol. 34, 2017, pp. 249–293.

[15] W. Song, W. Yao and Y. Xing. “Robust mixture regression model fitting by Laplace distribution”. Comp. Stat. Dat. Anal., vol. 71, 2014, pp. 128–137.

[16] W. Yao, Y. Wei and C. Yu. “Robust mixture regression using the t-distribution”. Comp. Stat. Dat. Anal., vol. 71, 2014, pp. 116–127.

[17] M. S. Yang, and Y. Nataliani. “Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters”. Patt. Recog., vol. 71, 2017, pp. 45–59.

[18] L. A. Zadeh. ‘Is there a need for fuzzy logic?”. Info. Scie., vol. 178, 2008, pp. 2751–2779.