Journal of Statistical Sciences, Spring and Summer, 2024
Vol. 18, No. 1, pp 193-212
DOI: 10.52547/jss.18.1.11

Robust Model-Based Clustering Using the Symmetric α-Stable Distribution for Measurement Error

Moradi, M., Zarei, S.
Department of Statistics, University of Kurdistan, Sanandaj, Iran.

Corresponding author: S. Zarei, sh.zarei@uok.ac.ir **Received:** 25/2/2024 **Revised:** 28/5/2024 **Accepted and Published Online:** 30/5/2024.

Introduction

Model-based clustering is the most widely used statistical clustering method. In this method, heterogeneous data are divided into homogeneous groups using inference based on mixture models. The presence of measurement errors in the data can reduce the quality of clustering and, for example, cause overfitting and produce spurious clusters. To solve this problem, model-based clustering assuming a normal distribution for measurement errors has been introduced. However, in practice, there are situations where errors are very large or small, known as outliers. For example, when recording people's income data, individuals often tend to report their income much higher or lower than the real amount, depending on their situation. Therefore, assuming normality in such cases is unrealistic and can reduce inference accuracy. In the case of outliers, mild outliers can be addressed using robust models or removed and analyzed separately. However, gross outliers are unpredictable and cannot be modeled using standard statistical distributions. Robust models based on heavy-tailed distributions, such as the t distribution, can be used to model and control mild outliers in data. Additionally, α-stable distributions can be employed to model gross outliers. For this reason, we use the symmetric α-stable distribution in the univariate case to model measurement errors, which can model normal, mild, and gross measurement errors depending on the value of α.

Material and Methods

In the literature on measurement error analysis, each observation consists of two latent parts: the true and measurement error values, which can be

random. When an outlier is observed in the observation, it may be due to measurement error or an outlier in the population. For modeling both cases simultaneously, a symmetric α-stable distribution is proposed as a replacement for the normal distribution for both measurement errors and real part of observations, and the model parameters are estimated using the EM algorithm and numerical methods. Furthermore, the optimal number of clusters can be determined by BIC.

Results and Discussion
The simulation results and real data analysis show that the proposed model performs better in cases where there are outlying observations in the data, mainly due to measurement errors, compared to MCLUST and MCLUST-ME methods. However, since the parameter α is estimated numerically, the proposed method requires more time, especially compared to the MCLUST algorithm, and may face convergence issues.

Conclusion
In practice, we can choose the appropriate clustering model through trial and error. For this purpose, one can fit the proposed model and other suitable models to the data and select the one that performs better based on performance evaluation metrics. Additionally, as one of the future objectives, we intend to extend this method to the multivariate case or use asymmetric α-stable distributions to model measurement error.

Keywords: Model-based clustering, α-Stable distribution, Measurement error, EM Algorithm.

Mathematics Subject Classification (2010): 60E07, 62H30.

مجله علوم آماری، بهار و تابستان ۱۴۰۳
جلد ۱۸، شماره ۱، ص ۱۹۳ -- ۲۱۲
DOI: 10.52547/jss.18.1.11
مقاله پژوهشی

خوشه‌بندی استوار مبتنی بر مدل با استفاده از توزیع α-پایدار متقارن برای خطای اندازه‌گیری

مژگان مرادی و شاهو زارعی

گروه آمار، دانشکده علوم پایه، دانشگاه کردستان

نویسنده مسئول : شاهو زارعی، sh.zarei@uok.ac.ir

تاریخ دریافت: ۱۴۰۲/۱۲/۶ تاریخ بازنگری: ۱۴۰۳/۳/۸ تاریخ پذیرش و انتشار: ۱۴۰۳/۳/۱۰

چکیده: خوشه‌بندی مبتنی بر مدل پرکاربردترین روش خوشه‌بندی آماری است که در آن داده‌های ناهمگن با استفاده از استنباط بر اساس مدل‌های آمیخته به گروه‌هایی همگن تقسیم می‌شوند. وجود خطای اندازه‌گیری در داده‌ها می‌تواند کیفیت خوشه‌بندی را کاهش و به عنوان مثال موجب بیش‌برازشی و تولید خوشه‌های جعلی شود. برای رفع این مشکل، خوشه‌بندی مبتنی بر مدل با فرض توزیع نرمال برای خطای اندازه‌گیری معرفی شده است. با وجود این، مقدارهای خیلی بزرگ یا خیلی کوچک (دورافتاده) از خطاهای اندازه‌گیری باعث عملکرد ضعیف روش‌های خوشه‌بندی موجود می‌شوند. برای رفع این مشکل و ساختن یک مدل استوار نسبت به حضور خطاهای اندازه‌گیری دورافتاده در داده‌ها، در این مقاله برای خطای اندازه‌گیری توزیع α-پایدار متقارن جایگزین توزیع نرمال می‌شود و با استفاده از الگوریتم EM و روش‌های عددی، پارامترهای مدل برآورد می‌شوند. با استفاده از شبیه‌سازی و تحلیل داده واقعی به مقایسه مدل جدید ارائه شده با روش خوشه‌بندی مبتنی بر مدل با روش $MCLUST$، در حالت‌های با و بدون خطای اندازه‌گیری پرداخته و کارایی مدل پیشنهادی برای خوشه‌بندی داده‌ها در حضور انواع خطاهای اندازه‌گیری دورافتاده، نشان داده می‌شود.

واژه‌های کلیدی: خوشه‌بندی مبتنی بر مدل، خطای اندازه‌گیری، توزیع α-پایدار، الگوریتم EM.

کد موضوع‌بندی ریاضی (۲۰۱۰): 60E07، 62H30.

۱ مقدمه

خوشه‌بندی مبتنی بر مدل یکی از پرکاربردترین الگوریتم‌های خوشه‌بندی موجود در آمار و ادبیات علوم داده و تعمیمی از روش K–میانگین است. در خوشه‌بندی به روش مبتنی بر مدل، فرض می‌شود دادهٔ $\boldsymbol{y}$ از توزیع آمیخته متناهی $g(\boldsymbol{y}) = \sum_{k=۱}^{G} \tau_k g_k(\boldsymbol{y}|\boldsymbol{\theta}_k)$ آمده باشد، که در آن $g_k(\cdot|\boldsymbol{\theta}_k)$ تابع چگالی احتمال جزء kام با پارامترهای $\boldsymbol{\theta}_k$ بوده و τ_k احتمال آن است که یک مشاهده از جزء kام آمده باشد، به‌طوری که $\tau_k \in (۰, ۱)$ و $\sum_{k=۱}^{G} \tau_k = ۱$. هر جزء معادل یک خوشه خواهد بود که رفتار آن توسط $g_k(\cdot|\boldsymbol{\theta}_k)$ بررسی یا مدل می‌شود. از این رو اصطلاح مدل آمیخته به جای توزیع آمیخته مرسوم‌تر است. معمولاً فرض می‌شود که تعداد مولفه‌ها متناهی است و مدل حاصل را مدل آمیخته متناهی می‌گویند. لازم به ذکر است متداول‌ترین مدل آمیخته برای خوشه‌بندی، مدل آمیخته گاوسی یا نرمال است. در این مدل فرض می‌شود که توزیع هر خوشه نرمال است و هدف از آن برآورد پارامترهای آن توزیع آمیخته به همراه متغیرهای پنهانی می‌باشد که به عنوان برچسب خوشه‌ها، در مدل معرفی شده‌اند. معمولاً خوشه‌بندی مبتنی بر مدل نرمال، بر اساس روش $MCLUST$ (فرالی و رافتری، ۲۰۰۳) انجام می‌شود که در بخش ۲ معرفی می‌شود.

اغلب الگوریتم‌های خوشه‌بندی موجود بر این اساس می‌باشند که داده‌ها فاقد خطای اندازه‌گیری هستند. دلیل اصلی این امر دوری از محاسبات پیچیده بررسی خطای اندازه‌گیری در داده‌هاست. با این حال، این یک فرضیه غیر واقعی و نادرست است چرا که وجود نقص در مراحل مختلف نظرسنجی‌ها یا در ثبت داده‌ها، می‌تواند موجب خطای اندازه‌گیری شود. به همین علت است که این موضوع اخیراً مورد توجه آمارشناسان قرار گرفته است. خطای اندازه‌گیری به تفاوت بین مقدار واقعی و مقدار اندازه‌گیری (یا مشاهده) شده از یک کمیت اشاره دارد. اصطلاح خطای اندازه‌گیری اولین بار توسط فولر (۲۰۰۹) در بررسی رابطه رگرسیونی بین ذرت با مقدار خالص نیتروژن موجود در خاک معرفی شد. این نوع خطا که می‌تواند تصادفی یا سیستماتیک باشد، ممکن است باعث کاهش دقت برآورد پارامترها و در نتیجه کاهش دقت استنباط شود. در خوشه‌بندی، خطای اندازه‌گیری می‌تواند خوشه‌های جعلی یا خوشه‌های مبهم ایجاد کند. همچنین می‌تواند بر شکل، فرم و پایداری آنها تأثیر بگذارد. برای بررسی بیشتر و دقیق‌تر تاثیر خطای اندازه‌گیری بر خوشه‌بندی می‌توان (پن‌کوسکا و اُبرسکی، ۲۰۲۰) را مطالعه کرد.

برای مدل کردن و بررسی خطای اندازه‌گیری در خوشه‌بندی معمولاً فرض می‌شود که این نوع خطا تصادفی و دارای توزیع نرمال است و برای مدل‌بندی آن روش $MCLUST - ME$ ارائه شده است (ژانگ و دی، ۲۰۲۰). حال آنکه در عمل موقعیت‌هایی وجود دارد که خطاها بسیار بزرگ یا کوچک یا به اصطلاح دورافتاده باشند. مثلا در ثبت کردن داده‌های درآمدی مردم اغلب تمایل دارند بستگی به وضعیت موجود، درآمد خود را خیلی بیشتر یا کمتر از مقدار واقعی بیان کنند. بنابراین در این حالت فرض نرمال بودن غیر واقعی و موجب کاهش دقت استنباط می‌شود. ریتر (۲۰۱۵) داده‌های دورافتاده را به دو نوع خفیف و شدید تقسیم کرد. اگر Q دامنه میان چارکی باشد، معمولاً داده‌ای که kQ ($k \geq ۱/۵$) از چارک اول کوچکتر یا از چارک سوم بزرگتر باشد، داده دورافتاده به حساب می‌آید. اگر $k \in [۱/۵, ۳)$ باشد، داده دورافتاده خفیف و اگر $k \geq ۳$ باشد، داده دور افتاده شدید است، (زارعی، ۱۴۰۰).

برای بررسی و کنترل داده‌های دور افتاده خفیف، می‌توان از مدل‌های استوار مبتنی بر توزیع‌های دم-کلفت[1] مانند توزیع t استفاده کرد. همچنین برای مدل‌بندی داده‌های دورافتاده شدید می‌توان از توزیع‌های α-پایدار[2] استفاده کرد، (نولن، ۲۰۲۰). به دلیل مشکلات محاسبه‌ای هنوز مدل آمیخته با خطای اندازه‌گیری دارای توزیع t بررسی نشده است. به همین خاطر ما در این مقاله در حالت یک متغیره توزیع α-پایدار متقارن را برای مدل کردن خطای اندازه‌گیری مورد استفاده قرار می‌دهیم که می‌تواند بستگی به مقدار α خطای اندازه‌گیری نرمال، خفیف و شدید را مدل کند.

در این مقاله، روش‌های $MCLUST$ و $MCLUST-ME$ در بخش ۲ معرفی می‌شوند. توزیع α-پایدار و خوشه‌بندی مبتنی بر مدل استوار با خطای اندازه‌گیری α-پایدار متقارن در بخش ۳ به طور کامل معرفی و نحوه برآورد پارامترهای مدل بیان می‌گردد. در بخش ۴ با مطالعه شبیه‌سازی، به ارزیابی و مقایسه مدل پیشنهادی با مدل‌های دارای خطای اندازه‌گیری و بدون خطای اندازه‌گیری می‌پردازیم. همچنین در بخش ۵ مدل پیشنهادی برای توصیف داده‌های واقعی مورد استفاده قرار گرفته و نهایتاً نتیجه‌گیری بیان می‌شود.

۲ روش‌های $MCLUST$ و $MCLUST-ME$

مهم‌ترین روش خوشه‌بندی در آمار، خوشه‌بندی مبتنی بر مدل است که در حالت چند متغیره معمولاً بر اساس روشی موسوم به $MCLUST$ انجام می‌شود (فرالی و رافتری، ۲۰۰۳). روش $MCLUST$ یک روش محاسبه برآورد پارامترهای مدل آمیخته گاوسی است که با قید گذاشتن روی ماتریس کواریانس، تعداد پارامترها را کاهش می‌دهد و سپس با توجه به ساختارهای مختلف به دست آمده برای ماتریس کواریانس خوشه‌بندی را انجام می‌دهد (بوویرون و همکاران، ۲۰۱۹). لازم به ذکر است برای انجام محاسبات مربوط به روش $MCLUST$ و انجام خوشه‌بندی مبتنی بر مدل می‌توان از بسته mclust (اسکروکا و همکاران، ۲۰۱۶) در نرم‌افزار R استفاده کرد. روش $MCLUST-ME$ در حقیقت تعمیم روش $MCLUST$ به حالتی است که داده‌ها دارای خطای اندازه‌گیری نرمال با ماتریس واریانس-کوواریانس معلوم باشند (ژانگ و دی، ۲۰۲۰). در این روش، محدودیت خاصی روی ماتریس کواریانس قرار نمی‌گیرد. همین امر باعث می‌شود وقتی تعداد خوشه‌ها و پارامترها زیاد باشد، سرعت همگرایی الگوریتم بسیار کم باشد. به عنوان مثال بر اساس مطالعه شبیه‌سازی، برای خوشه‌بندی یک مجموعه داده با ۲متغیر و ۱۰۰۰ تکرار به ۲ خوشه، ۱۹ دقیقه و خوشه‌بندی همین مجموعه داده به ۶ خوشه، نزدیک ۲۳ ساعت زمان لازم است (ژانگ و دی، ۲۰۲۰). در این مدل، فرض می‌شود که مشاهدات از دو قسمت $\boldsymbol{w}$ که نمایانگر بخش واقعی و $\boldsymbol{\epsilon}$ که نماینده قسمت مربوط به خطای اندازه‌گیری هستند، تشکیل شده باشند. ساختار آماری این مدل به‌صورت

$$\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{\epsilon}, \quad \boldsymbol{w} \mid k \sim N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \boldsymbol{\epsilon} \sim N_d(\circ, \boldsymbol{\Lambda}), \tag{۱}$$

[1] Heavy-tailed
[2] Stable

است، که در آن $\boldsymbol{w}$ و $\boldsymbol{\epsilon}$ هر دو دارای توزیع نرمال d متغیره و از هم مستقل هستند. همچنین $\boldsymbol{\mu}_k$ و $\boldsymbol{\Sigma}_k$ به ترتیب پارامترهای میانگین و کوواریانس مجهول $\boldsymbol{w}$ در خوشه kام می‌باشند و $\boldsymbol{\Lambda}$ ماتریس واریانس–کوواریانس معلوم خطای اندازه‌گیری $\boldsymbol{\epsilon}$ است. بنابراین $\boldsymbol{y}$ دارای توزیع آمیخته

$$g(\boldsymbol{y}) = \sum_{k=1}^{G} \frac{\tau_k}{\sqrt{\det\left[2\pi\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Lambda}\right)\right]}} e^{-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}_k\right)^T \left(\boldsymbol{\Sigma}_k+\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}_k\right)} \qquad (۲)$$

است. فرض معلوم بودن $\boldsymbol{\Lambda}$ یک فرض محدودکننده است. در عمل وقتی امکان تکرار اندازه‌گیری باشد، با استفاده از خودگردان‌سازی[1] (ژانگ و دی، ۲۰۲۰) یا با استفاده از ماتریس کواریانس نمونه‌ای می‌توان $\boldsymbol{\Lambda}$ را برآورد یا تقریب زد. همان‌طور که بیان شد روش $MCLUST-ME$ در حقیقت یک مدل آمیخته متناهی است. بنابراین برای برآورد پارامترها همانند روش $MCLUST$ از الگوریتم بیشینه‌سازی امید ریاضی (EM) (دمپستر و همکاران، ۱۹۷۷) استفاده می‌شود. الگوریتم EM یک روش تکرار شونده است که به دنبال یافتن براوردی با بیشترین درستنمایی برای پارامترهای یک توزیع پارامتری است. این الگوریتم روش متداول برای حالت‌هایی است که برخی از متغیرهای تصادفی پنهان هستند. الگوریتم EM به دنبال یافتن ماکسیمم موضعی بوده و در هر تکرار شامل دو مرحله است. مرحلهٔ اول یا مرحلهٔ E که تابع درستنمایی به‌وسیلهٔ امید ریاضی تقریب داده می‌شود و مرحلهٔ M که برآورد روش درستنمایی ماکسیمم پارامترها محاسبه می‌شود. ژانگ و دی (۲۰۲۰) با تشکیل تابع درستنمایی کامل، نشان دادند که براوردگرهای τ_k و $\boldsymbol{\mu}_k$ در تکرار $(t+1)$ام از رابطه‌های

$$\begin{aligned}
\hat{\tau}_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{k} e_{ik}^{(t)} \\
\hat{\boldsymbol{\mu}}_k^{(t+1)} &= \Big[\sum_{i=1}^{n} e_{ik}^{(t)}\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Lambda}\right)^{-1}\Big]^{-1} \sum_{i=1}^{n} e_{ik}^{(t)}\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{y}_i,
\end{aligned} \qquad (۳)$$

به‌دست می‌آیند، که در آن برای $k = 1,\ldots,G$ و $i = 1,\ldots,n$ داریم

$$e_{ik}^{(t)} = E_{\boldsymbol{\theta}^{(t)}}(Z_{ik}|\boldsymbol{y}_i) = \frac{\tau_k^{(t)} g_k\left(\boldsymbol{y}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}, \boldsymbol{\Lambda}\right)}{\sum_{j=1}^{G} \tau_j^{(t)} g_j\left(\boldsymbol{y}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}, \boldsymbol{\Lambda}\right)}.$$

[1] Bootstrapping

همچنین برای برآورد $\boldsymbol{\Sigma}_k$ برای $k = ۱, \cdots, G$ باید معادله

$$\frac{\partial \ell_C}{\partial \boldsymbol{\Sigma}_k} = \frac{۱}{۲}\sum_{i=۱}^{n} e_{ik}^{(t)}(\boldsymbol{\Sigma}_k + \boldsymbol{\Lambda})^{-۱}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)(\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T(\boldsymbol{\Sigma}_k + \boldsymbol{\Lambda})^{-۱} - \frac{۱}{۲}\sum_{i=۱}^{n} e_{ik}^{(t)}(\boldsymbol{\Sigma}_k + \boldsymbol{\Lambda})^{-۱} = ۰$$

حل شود، که فاقد جواب تحلیلی است. برای به‌دست آوردن برآورد $\boldsymbol{\Sigma}_k$ در هر تکرار، از روش‌های عددی استفاده می‌شود. برای این منظور از روش شبه–نیوتونی حافظه–محدود $BFGS$ در تابع $optim$ در نرم افزار R استفاده می‌شود. پس از محاسبه $\hat{\boldsymbol{\Sigma}}_k$ با جایگذاری آن در رابطه (۳) برآورد $\boldsymbol{\mu}_k$ به‌دست می‌آید. الگوریتم با مقدارهای اولیه شروع کرده و در هر تکرار، پارامترها به ترتیبی که بیان شد (تا همگرا شدن الگوریتم) به‌هنگام می‌شوند.

۳ خوشه‌بندی مبتنی بر مدل استوار با خطای اندازه‌گیری α–پایدار متقارن

توزیع‌های پایدار (α–پایدار) یک خانواده غنی از توزیع‌های آماری هستند. این توزیع‌ها همزمان امکان بررسی دم کلفتی و چولگی در داده‌ها را فراهم می‌کنند (نولن، ۲۰۲۰). متغیر تصادفی یک متغیره α–پایدار عضو خانوادهٔ توزیع‌های چهار پارامتری است. شاخص پایداری یا شاخص دم $\alpha \in (۰, ۲]$، پارامتر چولگی $\beta \in [-۱, ۱]$، پارامتر مقیاس $\gamma > ۰$ و پارامتر مکان $\delta \in \mathbb{R}$ چهار پارامتر این توزیع هستند. در این مقاله اگر Y متغیر تصادفی یک متغیره α–پایدار با پارامترهای مذکور باشد، به‌صورت $Y \sim S(\alpha, \beta, \gamma, \delta)$ و چگالی آن با $f_S(y; \alpha, \beta, \gamma, \delta)$ نمایش داده می‌شوند. توزیع α–پایدار تعمیم توزیع نرمال است. پارامتر α رفتار دم توزیع را کنترل می‌کند و وقتی‌که $\alpha \to ۰$ توزیع دم کلفت‌تر است. برای $\alpha < ۲$ واریانس توزیع بی‌نهایت بوده و وجود ندارد، (نولن، ۲۰۲۰). این ویژگی باعث می‌شود که این توزیع برای بررسی مدل‌هایی با داده‌های دورافتاده شدید و به‌دست آوردن یک مدل استوار مناسب باشد. تابع چگالی متغیرهای تصادفی α–پایدار به جز چند مورد خاص، صورت تحلیلی ندارد. بنابراین برای بررسی خواص این توزیع‌ها از تابع مشخصه استفاده می‌شود. متغیر تصادفی Y دارای توزیع α–پایدار یک متغیره است اگر تابع مشخصهٔ آن به‌صورت

$$\varphi_Y(u) = \begin{cases} \exp\big(-\gamma^\alpha |u|^\alpha \big(۱ - i\beta \tan(\frac{\pi\alpha}{۲})\text{sign}(u)\big) + i\delta u\big), & \alpha \neq ۱, \\ \exp\big(-\gamma|u|\big(۱ + i\frac{۲}{\pi}\beta\ \text{sign}(u)\log|u|\big) + i\delta u\big), & \alpha = ۱, \end{cases}$$

باشد که در آن $\text{sign}(\cdot)$ تابع علامت است. تابع چگالی احتمال توزیع‌های α–پایدار با گرفتن تبدیل فوریه معکوس تابع مشخصه به صورت

$$f_s(y;\alpha,\beta,\gamma,\delta) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \varphi_Y(u)e^{-iyu}du.$$

به‌دست می‌آید (سالاس گونزالس و همکاران، ۲۰۰۹). لازم به ذکر است که اگر $\beta = 0$ باشد، توزیع را α–پایدار متقارن در اطراف δ می‌نامند و با $S\alpha S$ نمایش می‌دهیم. خانواده α–پایدار مثبت نیز زمانی که $\beta = 1$ و $\alpha < 1$ باشد، به‌دست می‌آید. همچنین توزیع نرمال با میانگین δ و واریانس $2\gamma^2$ حالت خاصی از توزیع α–پایدار متقارن به‌صورت $S(2, 0, \gamma, \delta)$ است (نولن، ۲۰۲۰).

هنگامی‌که در مشاهدات داده‌ای دور افتاده ملاحظه می‌شود، این داده دورافتاده ممکن است به علت خطای اندازه‌گیری باشد یا اینکه واقعا یک داده دورافتاده در نمونه باشد. برای مدل‌بندی همزمان این دو حالت و به‌دست آوردن یک مدل استوار در حضور مشاهدات دورافتاده، در معادلات (۱) فرض می‌کنیم

$$w|k \sim S(\alpha_k, 0, \gamma_{wk}, \delta_{wk}), \quad \epsilon|k \sim S(\alpha_k, 0, \gamma_e, 0).$$

برای شناساپذیری[1] مدل، γ_e معلوم فرض می‌شود. بر طبق خاصیت مجموع دو متغیر تصادفی α–پایدار (ساموردنیتسکی و تاکو، ۱۹۹۴) توزیع حاشیه‌ای y در خوشه kام به صورت

$$y|k \sim S(\alpha_k, 0, (\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{1}{\alpha_k}}, \delta_{wk}). \qquad (۴)$$

است. چون تابع‌های چگالی توزیع‌های α–پایدار در حالت کلی دارای فرم بسته نیستند، برآورد کردن پارامترها در مدل‌های شامل این توزیع‌ها کار سختی است. بنابراین برای استفاده از الگوریتم EM در مدل‌های آمیخته شامل این توزیع‌ها از خاصیت مقیاس آمیخته نرمال‌ها[2] استفاده می‌شود.

تعریف ۱ (خاصیت مقیاس آمیخته نرمال‌ها برای $S\alpha S$). فرض کنید که Z متغیر تصادفی نرمال با میانگین صفر و واریانس $2\gamma^2$ باشد و P متغیر تصادفی الفا پایدار مثبت، در این صورت متغیر تصادفی $\mu + \sqrt{P}Z$ دارای توزیع $S(\alpha, 0, \gamma, \mu)$ است.

همانند زارعی و محمدپور (۲۰۲۰) در این مدل جدید برای $i = 1, \cdots, n$ داده‌های کامل به صورت (y_i, p_i, z_i) است که y_iها مشاهدات و p_i, z_i متغیرهای پنهان را تشکیل می‌دهند. بنابراین تابع چگالی احتمال

[1] Identifiability
[2] Scale mixtures of normals

توام به صورت

$$\prod_{i=1}^{n} f(y_i, p_i, \boldsymbol{z}_i) = \prod_{i=1}^{n}\prod_{k=1}^{G} \left\{ f(z_{ik}) f(p_i|z_{ik}) f(y_i|p_i, z_{ik}) \right\}^{z_{ik}} \qquad (۵)$$

خواهد شد، که در آن $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})^T$ به‌طوری‌که

$$z_{ik} = \begin{cases} ۱ & \text{اگر } y_i \text{ متعلق به خوشه } k \text{ باشد} \\ ۰ & \text{سایر جاها.} \end{cases}$$

با توجه به تعریف ۱ و روابط (۴) و (۵) لگاریتم تابع درستنمایی داده‌های کامل به صورت

$$\begin{aligned} \ell_c(\theta) = C &+ \sum_{k=1}^{G}\sum_{i=1}^{n} \left[z_{ik} \log(\tau_k) + z_{ik} \log(f(p_i|z_{ik})) \right] \\ &+ \sum_{k=1}^{G}\sum_{i=1}^{n} \left[z_{ik} \log\left(\phi(y_i; \delta_{wk}, ۲p_i(\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}) \right) \right], \end{aligned} \qquad (۶)$$

است، که در آن ثابت C به پارامترهای مدل بستگی ندارد. در این حالت نیز برای برآورد پارامترها از الگوریتم EM استفاده می‌کنیم.

مرحله E: در این مرحله امید ریاضی شرطی $E(l_c(\theta)|y_i)$ را محاسبه می‌کنیم. بنابراین صرف نظر از ضرایب ثابت باید در تکرار tام، امید ریاضی‌های شرطی $e_{1ik}^{(t)} = E_{\boldsymbol{\Psi}^{(t)}}[Z_{ik}|y_i]$ و $e_{pik}^{(t)} = E_{\boldsymbol{\Psi}^{(t)}}[\frac{1}{P_i}|y_i]$ را محاسبه کنیم، که در آن $\boldsymbol{\Psi}^{(t)}$ بردار پارامترهای مدل است. برای محاسبه $e_{1ik}^{(t)}$ در حالت کلی و صرف نظر از مرحله تکرار الگوریتم، داریم

$$\begin{aligned} e_{1ik} &= E_{\boldsymbol{\Psi}^{(t)}}[Z_{ik}|y_i] \\ &= \frac{P(y_i, Z_{ik} = ۱)}{f(y_i)} \\ &= \frac{\tau_k P(y_i|Z_{ik} = ۱) P(Z_{ik} = ۱)}{\sum_{k=1}^{G} \tau_k P(y_i|Z_{ik} = ۱) P(Z_{ik} = ۱)} \\ &= \frac{\tau_k f_S(y_i; \alpha_k, ۰, (\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{1}{\alpha_k}}, \delta_{wk})}{\sum_{k=1}^{G} \tau_k f_S(y_i; \alpha_k, ۰, (\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{1}{\alpha_k}}, \delta_{wk})}. \end{aligned}$$

همچنین

$$e_{pik}^{(t)} = E_{\Psi^{(t)}}[\frac{۱}{P_i}|y_i] = \int_{p_i} \frac{۱}{p_i} f(p_i|y_i)dp_i \qquad (۷)$$

چون $f(p_i|y_i)$ دارای فرم بسته نیست، این انتگرال دارای جواب تحلیلی نیست، اما می‌توان آن را به روش‌های عددی حل یا به عبارتی تقریب زد. ما از روش انتگرال مونت کارلو (کنگ و همکاران، ۲۰۰۳) این انتگرال را محاسبه می‌کنیم. برای محاسبه (۷) برای راحتی کار از اندیس i صرف نظر می‌کنیم. چون $y \sim f_S(y|\alpha_k, ۰, ۲(\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}, \delta_{wk})$ است، بنابراین با توجه به تعریف ۱، $y \overset{d}{=} \delta_{wk} + \sqrt{P}Z$ که Z متغیر تصادفی گاوسی با میانگین صفر با واریانس $۲(\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}$ است و $P \sim S(\frac{\alpha_k}{۲}, ۱, (\cos(\frac{\pi\alpha_k}{۴}))^{\frac{۲}{\alpha_k}}, ۰)$ یک متغیر تصادفی پایدار مثبت است. برای محاسبه $E_۱ = E\Big(P^{-۱}|y; \alpha_k, ۲(\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}, \delta_{wk}\Big)$، باید $f(p|y) = \frac{f(y,p)}{f(y)}$ را محاسبه کنیم. از آنجایی که $\frac{f(p)f(y|p)}{\int_۰^\infty f(p)f(y|p)dp}$ و $y|P = p \sim N(\delta_{wk}, ۲p(\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}})$، داریم

$$E_۱ = \frac{\int_۰^\infty p^{-۱/۲-۱} f_P(p|\alpha_k) \exp\{\frac{-(y-\delta_{wk})^۲}{۴p(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}}\}dp}{\int_۰^\infty p^{-۱/۲} f_P(p|\alpha_k) \exp\{\frac{-(y-\delta_{wk})^۲}{۴p(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}}\}dp}.$$

برای تقریب $E_۱$، از روش مونت کارلو با تولید B نمونه از تابع چگالی P و محاسبه عناصر زیر انتگرال استفاده می‌کنیم. اگر $p_۱^{mc}, \ldots, p_B^{mc}$ یک نمونه تصادفی از $f_P(p|\alpha_k)$ باشد، آنگاه مقدار تقریبی $E_۱$ به صورت زیر است:
برای تقریب $E_۱$، از روش مونت کارلو با تولید B نمونه از تابع چگالی P و محاسبه عناصر زیر انتگرال استفاده می‌کنیم. اگر $p_۱, \ldots, p_B$ یک نمونه تصادفی از $f_P(p|\alpha_k)$ باشد، آنگاه مقدار تقریبی $E_۱$ برابر

$$\frac{\sum_{b=۱}^B p_b^{mc(\frac{-۱}{۲}-۱)} \exp\{\frac{-(y-\delta_{wk})^۲}{۴p_b^{mc}(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}}\}}{\sum_{b=۱}^B p_b^{mc(\frac{-۱}{۲})} \exp\{\frac{-(y-\delta_{wk})^۲}{۴p_b^{mc}(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}}\}}.$$

است. $B = ۲۰۰۰$ انتخاب شده و $e_{pik}^{(t)}$ در تکرار tام برای $i = ۱, \ldots, N$ و $g = ۱, \ldots, G$ به‌هنگام می‌شود.
مرحله M: در این مرحله براوردگر ماکسیمم درستنمایی پارامترهای مدل بر اساس تابع زیر برآورد می‌شوند.

$$\ell_{ec}(\theta) = C + \sum_{k=۱}^G \sum_{i=۱}^n \big[e_{۱ik} \log(\tau_k) + e_{۱ik} E\big(\log(f(p_i|\alpha_k))|y_i\big) + e_{۱ik}\frac{-۱}{\alpha_k}\log(\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k}) + e_{۱ik}e_{pik}\frac{-(y_i - \delta_{wk})^۲}{۴(\gamma_{wk}^{\alpha_k} + \gamma_e^{\alpha_k})^{\frac{۲}{\alpha_k}}}\big]$$

برای به‌هنگام کردن پارامترهای مدل به راحتی می‌توان نشان داد

$$\hat{\tau}_k = \frac{1}{n}\sum_{i=1}^{n} e_{1ik}, \quad \hat{\delta}_{wk} = \frac{\sum_{i=1}^{n} e_{1ik}e_{pik}y_i}{\sum_{i=1}^{n} e_{1ik}e_{pik}}.$$

همچنین برای برآورد γ_{wk} چون $y_i|p_i, z_{ik} \sim N\big(\delta_{wk}, 2p_i\big(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k}\big)^{\frac{2}{\alpha_k}}\big)$، باید تابع

$$\sum_{i=1}^{n}\Big[e_{1ik}\frac{-1}{\alpha_k}\log(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k}) - e_{1ik}e_{pik}\cdot\frac{(y_i-\hat{\delta}_{wk})^2}{4\big(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k}\big)^{\frac{2}{\alpha_k}}}\Big]$$

نسبت به γ_{wk} ماکسیمم شود. پس از انجام محاسبات لازم براوردگر درستنمایی γ_{wk} به صورت

$$\hat{\gamma}_{wk} = \Big[-\gamma_e^{\alpha_k} + \Big(\frac{\sum_{i=1}^{n} e_{1ik}e_{pik}\big(y_i-\hat{\delta}_{wk}\big)^2}{2\sum_{i=1}^{n} e_{1ik}}\Big)^{\frac{\alpha_k}{2}}\Big]^{\frac{1}{\alpha_k}}. \qquad (۸)$$

به دست می‌آید. چنانچه مقدار $\hat{\gamma}_{wk}$ در رابطه (۸) در تکرارهای الگوریتم EM منفی شد آن را برای حافظ دامنه شدن صفر قرار می‌دهیم. برای برآورد کردن α_k، $k = 1, \cdots, K$ با توجه به (۶) باید تابع

$$\begin{aligned}\ell_{\alpha_k}(\theta) = &\sum_{i=1}^{n} e_{1ik}\big[E\big(\log\big(f(p_i;\alpha_k)\big)|y_i\big)\big] \\ &-\sum_{i=1}^{n}\Big[\frac{1}{\alpha_k}\log(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k}) + \frac{(y_i-\hat{\delta}_{wk})^2}{4\big(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k}\big)^{\frac{2}{\alpha_k}}}\Big]. \qquad (۹)\end{aligned}$$

نسبت به α_k ماکسیمم شود. چون تابع چگالی احتمال $f(p_i;\alpha_k)$ از نوع α-پایدار است و فرم بسته‌ای برای آن در حالت کلی وجود ندارد، تابع درستنمایی نیز دارای فرم بسته نیست و نمی‌توان به طور تحلیلی آن را حل کرد. بنابراین مشابه محاسبات انجام شده برای تقریب (۷) داریم

$$E\Big(\log\big(f(p_i;\alpha_k)\big)|y_i\Big) \approx \frac{\sum_{b=1}^{B}\log(f(p_{bi}^{mc};\alpha_k))p_{bi}^{mc(\frac{-1}{2})}\exp\{\frac{-(y_i-\delta_{wk})^2}{4p_{bi}^{mc}(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{2}{\alpha_k}}}\}}{\sum_{b=1}^{B}p_{bi}^{mc(\frac{-1}{2})}\exp\{\frac{-(y_i-\delta_{wk})^2}{4p_{bi}^{mc}(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{2}{\alpha_k}}}\}}.$$

بنابراین رابطه (۹) به صورت تقریبی

$$\ell_{\alpha_k}(\theta) \approx \sum_{i=1}^{n} e_{1ik}\Big[\frac{\sum_{b=1}^{B}\log(f(p_{bi}^{mc},\alpha_k))p_{bi}^{mc(\frac{-1}{2})}\exp\{\frac{-(y_i-\delta_{wk})^2}{4p_{bi}^{mc}(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{2}{\alpha_k}}}\}}{\sum_{b=1}^{B}p_{bi}^{mc(\frac{-1}{2})}\exp\{\frac{-(y_i-\delta_{wk})^2}{4p_{bi}^{mc}(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})^{\frac{2}{\alpha_k}}}\}}\Big]$$
$$-\sum_{i=1}^{n}\Big[\frac{1}{\alpha_k}\log(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k})+\frac{(y_i-\hat{\delta}_{wk})^2}{4\big(\gamma_{wk}^{\alpha_k}+\gamma_e^{\alpha_k}\big)^{\frac{2}{\alpha_k}}}\Big].$$

خواهد شد، که برای حل آن در تکرار tام، ابتدا برای هر i، نمونه‌ای به حجم B از توزیع α–پایدار مثبت تولید می‌کنیم. سپس α_k را مثلا ۱ قرار می‌دهیم. سپس با قرار دادن مقدار به‌هنگام شده‌ی سایر پارامترهای مجهول، مقدار سری را برای $\alpha_k = 1$ حساب می‌کنیم. سپس مقدار α_k را به اندازه مثلا ۰٫۱ افزایش داده و دوباره مجموع را محاسبه می‌کنیم. مقداری از α_k که بیشترین مقدار مجموع را باعث شود، مقدار برآورد شده برای α_k است. لازم به ذکر است که برای تولید اعداد تصادفی از توزیع α–پایدار مثبت، محاسبه مقدارهای تابع چگالی در یک یا چند نقطه و سایر محاسبات مربوط به توزیع‌های α–پایدار در نرم افزار R می‌توان از بسته STABLE، قابل دسترس در http://www.robustanalysis.com استفاده کرد.

۴ ارزیابی مدل $S\alpha SMEMBC$

در این بخش بر اساس چند مطالعه شبیه‌سازی، مدل استوار پیشنهادی را که با $S\alpha SMEMBC$ نمایش داده می‌شود، با روش خوشه‌بندی مبتنی بر مدل بدون خطای اندازه‌گیری با روش $MCLUST$ و روش خوشه‌بندی مبتنی بر مدل با روش $MCLUST$ و در حضور خطای اندازه‌گیری یعنی $MCLUST-ME$ (ژانگ و دی، ۲۰۲۰) برای اندازه‌های نمونه‌ای ۲۰۰ و ۴۰۰ مورد مقایسه قرار می‌دهیم. در عمل، برای انتخاب مدل مناسب و تعداد مؤلفه‌ها می‌توان از معیار اطلاع بیزی[1] (BIC)(شوارتز، ۱۹۷۸)

$$BIC = 2\log(L(\hat{\boldsymbol{\theta}})) - m\log(n),$$

استفاده کرد، که در آن $\hat{\boldsymbol{\theta}}$ برآورد ماکسیمم درستنمایی $\boldsymbol{\theta}$ ، $\log(L(\hat{\boldsymbol{\theta}}))$ لگاریتم مقدار ماکسیمم تابع درستنمایی مشاهده شده و m تعداد کل پارامترهای آزاد در مدل است و در مدل پیشنهادی $m = 4G-1$ است.

از آنجایی که داده‌ها ممکن است حاوی مقادیر دورافتاده باشند، برای تعیین مقادیر اولیه، از نتایج خوشه‌بندی حاصل از روش خوشه‌بندی $k-median$ با $k=G$ استفاده می‌کنیم. این بدان معناست که پس از تقسیم داده‌ها به G گروه بر اساس $k-median$، در هر گروه، مقادیر α_k، γ_{wk} و δ_{wk} را برای $k = 1,\cdots,G$ را با استفاده از

[1] Bayesian Information Criterion

بسته $STABLE$ برآورد می‌کنیم و به عنوان مقدار اولیه استفاده می‌کنیم. به عنوان یک قاعده کلی، الگوریتم زمانی متوقف می‌شود که تغییر نسبی لگاریتم درستنمایی داده‌های مشاهده شده یعنی $\frac{\log L(\psi^{(m+1)}) - \log L(\psi^{(m)})}{|\log L(\psi^{(m)})|}$ به مقدار آستانه مشخص شده (به عنوان مثال $\epsilon = ۱۰^{-۴}$) برسد، به این مرحله به عنوان زمان داغیدن[1] اشاره می‌کنیم.

شبیه‌سازی ۱

در این شبیه‌سازی توانایی مدل در برآورد پارامترها را در دو حالت با و بدون خطای اندازه‌گیری بررسی می‌کنیم. مقدارهای واقعی پارامترها به همراه میانگین و انحراف معیار مقدارهای برآورد شده در ۱۰۰ بار تکرار شبیه‌سازی برای اندازه نمونه $n = ۳۰۰$ در جدول ۱ آمده است. لازم به ذکر است مقدار پارامتر مقیاس برای جمله خطای اندازه‌گیری $\sqrt{۲}$ است. با توجه به نتایج جدول ۱، مثلاً وقتی که در خوشه اول مقدار واقعی پارامتر $\alpha = ۱/۴$ است،

جدول ۱. میانگین و انحراف معیار برآورد پارامترها ، با و بدون خطای اندازه‌گیری برای $n = ۳۰۰$.

			بدون خطای اندازه‌گیری		با خطای اندازه‌گیری	
خوشه	پارامتر	مقدار واقعی	میانگین	انحراف معیار	میانگین	انحراف معیار
اول	α	۱/۴	۱/۴۸۱۲	۰/۰۹۱۳	۱/۶۱۷۲	۰/۰۸۶۸
	τ	۰/۵	۰/۴۷۲۲	۰/۰۲۰۲	۰/۴۳۸۸	۰/۰۳۵۸
	γ	۰/۵۷	۰/۶۰۱۲	۰/۰۵۷۲	۰/۷۸۸۷	۰/۴۵۱۹
	δ	−۳	−۳/۰۰۲۰	۰/۰۸۰۹	−۲/۸۴۱۴	۰/۳۶۸۷
دوم	α	۱/۷۵	۱/۸۴۹۲	۰/۰۷۶۰	۱/۷۶۲۱	۰/۰۶۶۸
	τ	۰/۵	۰/۵۲۷۵	۰/۰۲۰۱	۰/۵۶۱۲	۰/۰۳۵۸
	γ	۱/۲	۱/۷۶۰۱	۰/۰۹۶۸	۱/۲۱۲۱	۰/۳۳۱۴
	δ	۳	۳/۱۲۱۴	۰/۱۵۵۸	۳/۱۹۱۲	۰/۵۰۲۴
شاخص رند تعمیم‌یافته			۰/۸۷۱۲	۰/۰۴۵۴	۰/۴۷۹۳	۰/۰۵۷۹
معیار اطلاع بیز			−۱۴۸۴/۱۲۱۳	۵۱/۵۵۰۱	−۱۸۳۸/۸۸۴۵	۳۷/۹۹۱۲

در حالت با و بدون خطای اندازه‌گیری، میانگین مقدارهای متوسط برآورد شده به ترتیب ۱/۶۱ و ۱/۴۸ به‌دست آمده است. یعنی در حالت بدون خطای اندازه‌گیری نسبت به حالت با خطای اندازه‌گیری، مقدار میانگین به‌دست آمده به مقدار واقعی $\alpha = ۱/۴$ نزدیک‌تر می‌باشد. انحراف معیار نیز در هر دو حالت زیاد نمی‌باشد که نشان می‌دهد مقدارهای برآورد شده در تکرارهای مختلف زیاد متفاوت نیستند. به عنوان نمونه دیگر، در خوشه دوم نیز مقدار واقعی وزن آمیخته $\tau = ۰/۵$ می‌باشد که در حالت بدون خطای اندازه‌گیری مقدار متوسط برآورد شده ۰/۵۲۸ و در حالت با خطای اندازه‌گیری نیز ۰/۵۶۱ به‌دست آمده است که در این حالت نیز مقدار برآورد شده در حالت بدون خطای اندازه‌گیری نسبت به حالت با خطای اندازه‌گیری به مقدار واقعی پارامتر نزدیک‌تر می‌باشد. انحراف معیار برآورد شده نیز در هر دو روش کم است. با بررسی مقدارهای واقعی پارامترهای دیگر در دو خوشه، مقایسه میانگین و انحراف معیار آنها

[1] Burning

در دو حالت بدون و با خطای اندازه‌گیری در کل به این نتیجه می‌رسیم که پارامترهای قسمت بدون خطای اندازه‌گیری دقیق‌تر از قسمت با خطای اندازه‌گیری برآورد شده‌اند.

برای بررسی تاثیر خطای اندازه‌گیری بر کیفیت نتایج خوشه‌بندی انجام شده، از شاخص رند تعمیم‌یافته[1] (هوبرت و ارابی، ۱۹۸۵) و معیار اطلاع بیزی استفاده می‌کنیم. قابل ذکر است مقدار شاخص رند تعمیم‌یافته هرچه به یک نزدیک‌تر باشد نمایانگر خوشه‌بندی مطلوب‌تر است. با توجه به جدول ۱، این شاخص در حالت بدون خطای اندازه‌گیری ۰٫۸۷ و در حالت با خطای اندازه‌گیری ۰٫۴۷ به‌دست آمده است. مقادیر شاخص رند تعمیم‌یافته به‌دست آمده در کل نشان می‌دهد که وجود خطای اندازه‌گیری باعث کاهش دقت خوشه‌بندی می‌گردد. با مقایسه معیار اطلاع بیزی محاسبه شده نیز می‌توان به تاثیر وجود خطای اندازه‌گیری، در کاهش دقت خوشه‌بندی پی برد.

شبیه‌سازی ۲

با فرض آنکه داده‌ها دارای دو خوشه هستند، برای تولید داده‌ها مدل $y_i = w_i + \epsilon_i$، $i = ۱, \ldots, n$ با شش سناریو در نظر گرفته شده است:

سناریوی اول و دوم: w_i و ϵ_i دارای توزیع α_k–پایدار متقارن با شاخص‌های پایداری متفاوت هستند.

سناریوی سوم: w_i و ϵ_i دارای توزیع نرمال هستند.

سناریوی چهارم: w_i دارای توزیع α_k–پایدار متقارن است و خطای اندازه‌گیری نداریم.

سناریوی پنجم: w_i دارای توزیع نرمال است و خطای اندازه‌گیری نداریم.

سناریوی ششم: w_i دارای توزیع نرمال و ϵ_i دارای توزیع t با ۳ درجه آزادی هستند.

تفاوت سناریوهای اول و دوم در مقدارهای شاخص پراکندگی است. بدین معنی که در سناریوی اول چون مقدارهای این شاخص‌ها کوچکتر است، احتمال تولید داده‌های دور افتاد شدید نسبت به خفیف بیشتر است. هدف از سناریوهای سوم و چهارم، ارزیابی مدل‌های مورد مقایسه در خوشه‌بندی داده‌های فاقد خطای اندازه‌گیری است. همچنین در سناریوی ششم، توانایی مدل‌های مورد مقایسه، در خوشه‌بندی داده‌های دارای خطای اندازه‌گیری، درحالتی که فرضیات مدل‌ها برقرار نباشد را مورد بررسی قرار می‌دهیم.

مقدارهای پارامترها برای سناریوهای مورد اشاره به صورت زیر است. مقدارهای پارامتر α در خوشه یک در حالت شدید و خفیف، به ترتیب ۱٫۱ و ۱٫۴۵ و این مقدارها در خوشه دوم ۱٫۲ و ۱٫۸ است. همچنین مقدار این پارامتر در حالت نرمال ۲ در نظر گرفته می‌شود. مقدارهای پارامترهای δ، γ و τ در خوشه اول به ترتیب ۳−، ۰٫۵۷ و ۰٫۴۵ و در خوشه دوم این مقدارها به ترتیب ۳، ۱٫۲ و ۰٫۵۵ است که برای وضعیت‌های مختلف ثابت فرض می‌شود. همچنین برای بررسی اثر اندازه نمونه، دو اندازه ۲۰۰ و ۴۰۰ را بررسی می‌کنیم. با بررسی نتایج به‌دست آمده برای شاخص رند تعمیم‌یافته، بر اساس ۱۰۰ بار تکرار شبیه‌سازی، در جدول ۲ به طور کلی می‌توان گفت که با افزایش اندازه نمونه، دقت خوشه‌بندی نیز بهتر شده است. الگوریتم پیشنهادی $S\alpha SMEMBC$ در حالتی که توزیع داده‌ها

[1] Adjusted Rand Index

جدول ۲. میانگین و انحراف معیار برآورد ها برای شاخص رند تعمیم‌یافته.

n	مدل	سناریوی اول	سناریوی دوم	سناریوی سوم	سناریوی چهارم	سناریوی پنجم	سناریوی ششم
۲۰۰	$S\alpha SMEMBC$	۰٫۳۵۶(۰٫۱۶۳)	۰٫۴۴۳(۰٫۲۴۹)	۰٫۱۹۲(۰٫۲۴۱)	۰٫۷۴۶(۰٫۰۶۶)	۰٫۸۴۷(۰٫۱۰۴)	۰٫۴۵۴(۰٫۱۴۴)
	$MCLUST-ME$	۰٫۰۰۰(۰٫۰۰۸)	۰٫۱۶۷(۰٫۲۴۹)	۰٫۵۴۶(۰٫۲۴۹)	۰٫۱۸۱(۰٫۲۶۵)	۰٫۹۳۵(۰٫۰۴۲)	۰٫۱۶۲(۰٫۱۲۹)
	$MCLUST$	−۰٫۰۰۱(۰٫۰۰۷)	۰٫۰۰۴(۰٫۰۰۷)	۰٫۰۹۱(۰٫۳۲۵)	۰٫۶۰۷(۰٫۱۳۱)	۰٫۹۳۸(۰٫۰۴۶)	۰٫۰۱۲(۰٫۰۹۵)
۴۰۰	$S\alpha SMEMBC$	۰٫۳۷۱(۰٫۱۳۴)	۰٫۵۸۰(۰٫۰۳۰)	۰٫۱۲۸(۰٫۰۰۴)	۰٫۷۵۹(۰٫۰۳۲)	۰٫۸۶۱(۰٫۵۰۴)	۰٫۵۱۸(۰٫۱۰۷)
	$MCLUST-ME$	۰٫۰۰۷(۰٫۰۰۹)	۰٫۲۶۲(۰٫۴۱۲)	۰٫۴۳۳(۰٫۲۴۱)	−۰٫۰۰۱(۰٫۰۰۵)	۰٫۹۵۶(۰٫۰۲۶)	۰٫۱۸۰(۰٫۱۱۴)
	$MCLUST$	۰٫۰۰۴(۰٫۰۰۹)	۰٫۰۰۵(۰٫۱۰۲)	۰٫۵۸۶(۰٫۰۴۹)	۰٫۵۹۷(۰٫۱۱۴)	۰٫۹۴۴(۰٫۰۳۸)	۰٫۰۱۲(۰٫۰۳۵)

و خطای اندازه‌گیری α–پایدار متقارن است، هم در حالت داده دورافتاده خفیف و هم شدید، بهتر از مدل‌های رقیب عمل می‌کند. همچنین در سناریوی چهارم نیز که توزیع داده‌ها α–پایدار متقارن بوده و فاقد خطای اندازه‌گیری، باز هم مدل پیشنهادی عملکرد بهتری دارد. لازم به ذکر است که در این سناریو نیز مشاهدات همراه با داده‌های دور افتاده ثبت شده‌اند و در عمل تفکیک اینکه مشاهده دور افتاده واقعی یا به علت خطای اندازه‌گیری است یا خیر، سخت می‌باشد. در سناریوی سوم که با مشاهدات تولید شده از توزیع نرمال روبرو هستیم، مدل‌های رقیب عملکرد بهتری از مدل پیشنهادی دارند. همچنین در سناریوی پنجم که با خوشه‌بندی داده‌های نرمال و بدون داده دورافتاده روبرو هستیم، کیفیت خوشه‌بندی نسبت به سایر سناریوها بهتر شده است. با وجود این در مدل پیشنهادی به علت اینکه پارامترها به روش عددی برآورد می‌شوند و دقت برآورد اندکی کاهش می‌یابد، از مدل‌های رقیب ضعیف‌تر عمل کرده است. همچنین در سناریوی ششم با وجود بدمشخص‌سازی[1]، مدل‌های خوشه‌بندی در حضور خطای اندازه‌گیری، کارایی بیشتری نسبت به مدل $MCLUST$ دارند.

۵ تحلیل داده‌های آنزیم

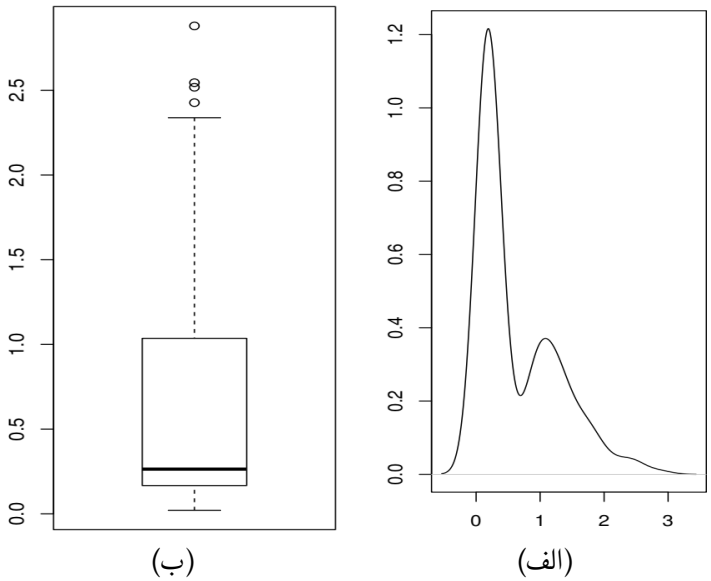
در این قسمت مجموعه داده آنزیم که در بسته mixAK (کومارک و کومارکوا، ۲۰۱۴) نرم‌افزار R وجود دارد را مورد بررسی قرار می‌دهیم. این داده‌ها از اطلاعات مربوط به آنزیم‌های خون ۲۴۵ نفر جمع‌آوری شده‌اند. بچتل و همکاران (۱۹۹۳) با برازش یک مدل آمیخته دو مولفه‌ای چوله به این داده‌ها، آنها را تحلیل کرده‌اند. همچنین زارعی و محمدپور (۲۰۲۰) با برازش یک مدل آمیخته دو مولفه‌ای α–پایدار متقارن به این داده‌ها آنها را مورد بررسی قرار داده‌اند. ابتدا با استفاده از BIC تعداد خوشه‌های بهینه را محاسبه می‌کنیم. برای این کار فرض می‌کنیم $G \in \{۱, ۲, ۳, ۴, ۵\}$ باشد و مقدار این شاخص را برای هر G محاسبه می‌کنیم. همان طور که از جدول ۳ مشخص است بیشترین مقدار BIC برای $G = ۲$ به‌دست آمده است و این نشان‌دهنده آن است که تعداد بهینه خوشه‌ها ۲ است. برای بررسی بیشتر در شکل ۱ نمودار تابع چگالی داده‌ها به همراه نمودار جعبه‌ای آنها رسم شده است. نمودار تابع چگالی چند مدی و در نمودار جعبه‌ای داده‌های دورافتاده وجود دارند که به ترتیب، نشان‌دهنده وجود ساختار خوشه‌ای در داده‌ها

[1] Misspecification

و همچنین احتمال وجود داده‌هایی با خطاهای اندازه‌گیری دورافتاده هستند. بنابراین این داده‌ها برای اهداف مقاله می‌توانند مناسب باشند.

جدول ۳. مقایسه مقدارهای BIC برای مدل $S\alpha SMEMBC$.

	BIC	$G=۱$	$G=۲$	$G=۳$	$G=۴$	$G=۵$
		−۶۰۴٫۳۱۹۸	−۵۵۹٫۱۵۰۴	−۵۷۴٫۶۹۱۷	−۵۸۶٫۸۹۴۵	−۵۹۴٫۱۶۱۳



شکل ۱. نمودار تابع چگالی احتمال داده‌های آنزیم (الف) و نمودار جعبه‌ای آنها (ب).

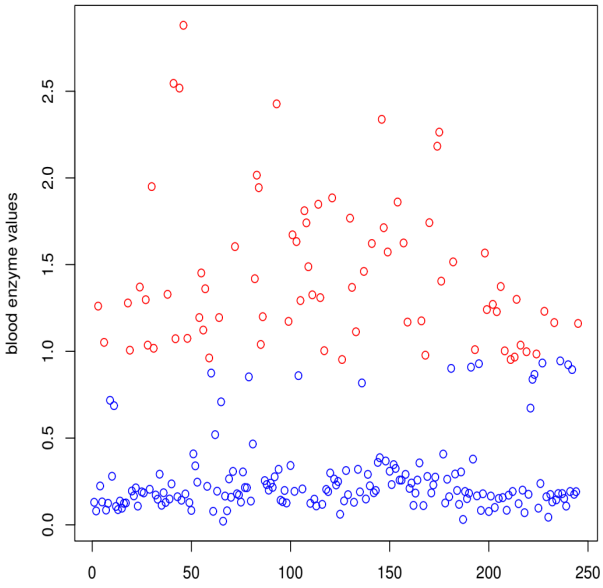
چون خوشه‌های واقعی داده‌ها معلوم نیستند، برای ارزیابی مدل‌ها از شاخص‌های دان[1](دان، ۱۹۷۴) و ضریب پهنای نیم‌رخ[2] (روسیو، ۱۹۸۷) استفاده می‌کنیم. لازم به ذکر هست هر چه مقدار این شاخص‌ها بیشتر باشد نشان‌دهنده خوشه‌بندی بهتر است. مقدارهای این ضریب‌ها برای $G=۲$ در جدول ۴ برای روش‌های خوشه‌بندی مختلف مورد مقایسه، آمده است. همچنین انحراف معیار داده‌ها تقسیم بر $\sqrt{۲}$ به عنوان مقدار γ_e تعیین شد.

با توجه به مقدارهای به‌دست آمده برای شاخص‌های ارزیابی در جدول ۴ می‌توان نتیجه گرفت که مدل پیشنهادی در خوشه‌بندی داده‌های آنزیم عملکرد بهتری از مدل‌های رقیب دارد. مقدارهای محاسبه شده بر اساس مدل پیشنهادی

[1] Dunn
[2] Silhouette

جدول ۴. مقدارهای ضریب پهنای نیم‌رخ و دان برای مدل‌های $S\alpha SMEMBC$، $MCLUST-ME$ و $MCLUST$ در خوشه بندی داده آنزیم.

نوع مدل آمیخته	ضریب پهنای نیم‌رخ	شاخص دان
$S\alpha SMEMBC$	۰٫۷۶۸۷	۰٫۰۶۸۹
$MCLUST-ME$	۰٫۶۴۴۶	۰٫۰۰۵۹
$MCLUST$	۰٫۷۴۳۷	۰٫۰۰۸۲



شکل ۲. خوشه‌بندی داده‌های آنزیم خون به دو گروه بر اساس مدل $S\alpha SMEMBC$.

برای وزن‌های آمیخته در خوشه‌های اول و دوم به ترتیب ۰٫۷۱ و ۰٫۲۹ است. مقدار شاخص‌های پایداری در خوشه‌های اول و دوم به ترتیب ۱٫۸۱ و ۱٫۸۳ برآورد شدند. همچنین مقدار پارامترهای مقیاس به همان ترتیب ۰٫۳۸ و ۰٫۴۶ محاسبه شدند. علاوه‌براین مقدارهای ۰٫۵۷ و ۰٫۹۶ به ترتیب در خوشه‌های اول و دوم برای شاخص‌های مکانی به‌دست آمدند. نتیجه خوشه‌بندی بر اساس مدل پیشنهادی در شکل ۲ نشان می‌دهد که مقدار آنزیم‌ها در گروه اول (نقاط آبی رنگ) از مقدار آنزیم‌ها در گروه دوم (نقاط قرمز رنگ) کمتر است.

بحث و نتیجه‌گیری

در این مقاله برای اولین بار خوشه‌بندی استوار مبتنی بر مدل، برای حالتی‌که مشاهدات دارای خطای اندازه‌گیری دورافتاده هستند، معرفی شد. برای این کار فرض شد که هم مشاهدات و هم خطاهای اندازه‌گیری دارای توزیع α–پایدار متقارن هستند و پارامترهای مدل به روش ماکسیمم درستنمایی و با استفاده از الگوریتم EM و روش‌های عددی برآورد شدند. همان‌طور که در تحلیل داده‌های آنزیم نشان داده شد، در عمل با استفاده از BIC می‌توان تعداد خوشه‌های بهینه را تعیین کرد. نتایج شبیه‌سازی‌ها و تحلیل داده‌های واقعی نشان می‌دهد که در حالتی که در داده‌ها مشاهدات دور افتاده مخصوصا به علت خطای اندازه‌گیری وجود دارد، مدل پیشنهادی دارای عملکرد بهتری نسبت به $MCLUST$ و $MCLUST - ME$ می‌باشد. اما چون پارامتر α با روش عددی برآورد می‌شود، روش پیشنهادی مخصوصاً نسبت به الگوریتم $MCLUST$ نیازمند زمان بیشتری است و ممکن است با مشکل همگرایی روبه‌رو شود. در عمل می‌توان به روش آزمایش و خطا مدل مناسب خوشه‌بندی را انتخاب کرد. برای این منظور می‌توان مدل پیشنهادی و مدل‌های مناسب دیگر را به داده‌ها برازش داد و هر کدام بر اساس شاخص‌های ارزیابی عملکرد بهتری داشت را انتخاب کرد. همچنین به عنوان یکی از اهداف آینده در نظر داریم این روش را به حالت چند متغیره گسترش دهیم. تمام کدهای مقاله در نرم‌افزار R نوشته شده‌اند و خواننده علاقه‌مند می‌تواند با مکاتبه با نویسنده مسئول مقاله به آنها دسترسی داشته باشد.

تقدیر و تشکر

نویسندگان مقاله از پیشنهادها و نظرهای ارزشمند داوران، سردبیر و ویراستار محترم مجله که باعث اصلاحات سازنده در محتوا و افزایش کیفیت مقاله شده است، کمال تشکر و قدردانی را دارند.

مراجع

تیموری، م. (۱۳۹۹)، براوردگر ماکسیمم درستنمایی توزیع آلفا–پایدار، مجله علوم آماری، ۱۴، ۷۳–۹۴.

زارعی، ش. (۱۴۰۰)، برآورد کوچک ناحیه‌ای بیز تجربی استوار با توزیع α–پایدار، مجله علوم آماری، ۲(۱۵)، ۴۶۳–۴۸۰.

Bechtel, Y. C., Bonaiti-Pellie, C., Poisson, N., Magnette, J., and Bechtel, P. R. (1993), A Population and Family Study N-Acetyltransferase Using Caffeine Urinary Metabolites. *Clinical Pharmacology & Therapeutics*, **54**(2), 134–141.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019), Model-Based Clustering and Classification for Data Science: with Applications in R. *Cambridge University Press*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

Dunn, J. C. (1974), Well-Separated Clusters and Optimal Fuzzy Partitions, *Journal of Cybernetics*, **4**(1), 95–104.

Fraley, C., and Raftery, A. E. (2003), Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST. *Journal of classification*, **20**(2), 263-286.

Fuller, W. A. (2009), *Measurement Error Models*, John Wiley & Sons.

Hubert, L., and Arabie, P. (1985), Comparing Partitions. *Journal of classification*, 2, 193–218.

Komárek, A., and Komárková, L. (2014), Capabilities of R Package mixAK for Clustering Based on Multivariate Continuous and Discrete Longitudinal Data. *Journal of Statistical Software*, **59**(12), 1–38.

Kong, A., McCullagh, P., Meng, X. L., Nicolae, D., and Tan, Z. (2009), A Theory of Statistical Models for Monte Carlo Integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(3), 585–604.

Nolan, J. P. (2020), *Stable Distributions: Models for Heavy-Tailed Data.* Springer Cham.

Pankowska, P., and Oberski, D. L. (2020), The effect of Measurement Error on Clustering Algorithms. *arXiv preprint arXiv*, :2005.11743.

Ritter, G. (2015), Robust Cluster Analysis and Variable Selection, *Vol. 137 of Chapman & Hall/CRC Monographs on Statistics & Applied Probability*, CRC Press.

Rousseeuw, P. J. (1987), Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Salas-Gonzalez, D., Kuruoglu, E. E., and Ruiz, D. P. (2009), Finite Mixture of α-Stable Distributions. *Digital Signal Processing* , 250–264.

Samorodnitsky, G. and Taqqu, M. S. (1994), *Stable Non-Gaussian Random Processes*, Chapman and Hall, New York.

Schwarz, G. (1978), Estimating the Dimension of a Model. *The Annals of Statistics*, 461–464.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016), mclust 5: Mlustering, Classification and Ddensity Estimation using Gaussian Finite Mixture Models. *The R Journal*, **8**(1), 205–233.

Zarei, S., and Mohammdpour, A. (2020), Pseudo-Stochastic EM for sub-Gaussian α-Stable Mixture Models. *Digital Signal Processing.* `doi.org/10.1016/j.dsp.2020.102671`. **99** 102671.

Zhang, W., and Di, Y. (2020), Model-Based Clustering with Measurement or Estimation Errors, *Genes*, **11**(2), 185–209.