Journal of Statistical Sciences, Spring and Summer, 2022
Vol. 16, No. 1, pp 63-89
DOI: 10.29252/jss.16.1.63

Clustering Based on Nonparanormal Graphical Mixture Models

Haji Aghabozorgi, H., Eskandari, F.
Department of Statistics, Allameh Tabatabaei University, Tehran, Iran.

Corresponding author: F. Eskandari, askandari@atu.ac.ir

Received: 3 December 2021 **Revised:** 25 July 2022 **Accepted and Published Online:** 1 August 2022.

Introduction

Graphical mixture models provide a powerful tool to visually depict the conditional independence relationships between high-dimensional heterogeneous data. In the study of graphical mixture models, the distribution of the mixture components is mostly considered multivariate normal with different covariance matrices. The resulting model is the Gaussian graphical mixture model (GGMM). The nonparanormal graphical mixture model (NGMM) has been introduced by replacing the normal assumption with a semiparametric Gaussian copula, which extends the nonparanormal graphical model and mixture models. This study proposes clustering based on NGMM under two forms of ℓ_1 penalty functions. Its performance is compared with clustering based on GGMM, in terms of cluster reconstruction and parameters estimation.

Material and Methods

The clustering based on NGMM is performed via a penalized EM algorithm under conventional and unconventional forms of ℓ_1 penalty functions (denoted by $NGMM_0$ and $NGMM_1$, respectively) and its performance over Gaussian and non-Gaussian simulated data sets are compared with the Gaussian ones (represented by $GGMM_0$ and $GGMM_1$, respectively). Along with the conventional ℓ_1 penalty, an alternative, unconventional penalty term is considered, which depends on the mixture proportions. Thus, the choice of mixture model distribution (Gaussian or nonparanormal) along with the choice of penalty function has emerged as the primary key of comparison. To

better compare the studied methods in terms of robustness against outliers, we considered deterministic and random contamination mechanisms. The proposed methodology is applied to Wisconsin diagnostic breast cancer data set to diagnose malignant or benign cancer patients.

Results and Discussion
The results of the simulation study on normal and nonparanormal datasets in ideal and noisy settings, as well as the application of breast cancer data set, showed that clustering approaches based on NGMM ($NGMM_0$ and $NGMM_1$) are more efficient and robust in the recovery of true cluster assignments than the clustering based on GGMM ($GGMM_0$ and $GGMM_1$), whereas, the unconventional PMLEs ($GGMM_1$ and $NGMM_1$) are more efficient in estimating the elements of precision matrices than the conventional PMLEs ($GGMM_0$ and $NGMM_0$).

Conclusion
The performance of clustering methods depends on the choice of penalty function and model selection, such that the combination of the nonparanormal graphical mixture model and the penalty term depending on the mixing proportions ($NGMM_1$) is more accurate than Gaussian ones in terms of cluster reconstruction and parameters estimation.

Keywords: Clustering, Graphical mixture models, Nonparanormal distribution, Penalized log-likelihood.
Mathematics Subject Classification (2010): 62H30, 62H20.

مجله علوم آماری، بهار و تابستان ۱۴۰۱
جلد ۱۶، شماره ۱، ص ۶۳ – ۸۹
DOI: 10.29252/jss.16.1.63
مقاله پژوهشی

خوشه‌بندی مبتنی بر مدل‌های آمیخته گرافی نرمال ناپارامتری

حمید حاجی‌آقابزرگی، فرزاد اسکندری

گروه آمار، دانشگاه علامه طباطبایی (ره)

نویسنده مسئول : فرزاد اسکندری، askandari@atu.ac.ir

تاریخ دریافت: ۱۴۰۰/۰۹/۱۲ تاریخ بازنگری: ۱۴۰۱/۰۵/۰۳ تاریخ پذیرش و انتشار: ۱۴۰۱/۰۵/۱۰

چکیده: مدل‌های آمیخته گرافی، ابزاری قدرتمند برای نمایش دیداری روابط استقلال شرطی بین داده‌های ناهمگن بالابُعد فراهم کرده است. در مطالعه این مدل‌ها، اغلب توزیع مولفه‌های آمیخته، نرمال چندمتغیره با ماتریس‌های کواریانس متفاوت در نظر گرفته شده که مدل حاصل، به مدل آمیخته گرافی گاوسی معروف است. با جای‌گزین کردن فرض محدودکننده نرمال با یک مفصل نیمه‌پارامتری نرمال، مدل آمیخته گرافی نرمال ناپارامتری معرفی شده که هم مدل گرافی نرمال ناپارامتری و هم مدل‌های آمیخته را تعمیم داده است. در این مطالعه، خوشه‌بندی مبتنی بر مدل آمیخته گرافی نرمال ناپارامتری با دو فرم تابع تاوان ℓ_1 (متعارف و نامتعارف) پیشنهاد شده است و عملکرد آن با روش خوشه‌بندی مبتنی بر مدل آمیخته گرافی گاوسی مقایسه شده است. نتایج مطالعه شبیه‌سازی روی داده‌های نرمال و غیرنرمال، در حضور و عدم حضور داده‌های دورافتاده و همچنین نتایج کاربردی روی داده‌های سرطان سینه نشان داد که ترکیب مدل آمیخته گرافی نرمال ناپارامتری با تابع تاوان وابسته به نسبت‌های آمیخته، از نظر بازسازی خوشه‌ها و برآورد پارامترهای مدل، نسبت به سایر روش‌های خوشه‌بندی مبتنی بر مدل از دقت بالاتری برخوردار است.

واژه‌های کلیدی: توزیع نرمال ناپارامتری، خوشه‌بندی، لگاریتم درستنمایی تاوانیده، مدل‌های آمیخته گرافی.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H20، 62H30.

۱ مقدمه

مدل‌های گرافی بی‌سو [۱] برای توصیف روابط وابستگی یا استقلال شرطی بین متغیرها در داده‌های بالابُعد [۲] به‌کار می‌رود. در این مدل‌ها، تعیین توزیع توام بردار تصادفی $\boldsymbol{X} = (X_1, \ldots, X_p)$ متناظر با راس‌های یک گراف G، چالشی مهم است. وقتی متغیرهای تصادفی متناظر با راس‌ها پیوسته باشند، توزیع آن‌ها اغلب به‌عنوان توزیع نرمال چند متغیره در نظر گرفته می‌شود ($\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$) و مدل حاصل، به مدل گرافی گاوسی [۳] (GGM) تعبیر می‌شود. در این حالت، می‌توان ماتریس دقت تُنُک [۴] $\Omega = \Sigma^{-1}$ را به‌طور مستقیم به یک مدل گرافی گاوسی تعبیر کرد. بنابراین، مساله برآورد مدل گرافی گاوسی به برآوردیابی ماتریس دقت محدود می‌شود. الگوریتم‌های بسیاری نظیر ماکسیمم درستنمایی ℓ_1-منظم شده [۵]، انتخاب همسایگی [۶] و الگوریتم لاسوی گرافی [۷]، برای برآورد ماتریس دقت در مدل گرافی گاوسی تعمیم داده شده است (مینشاوسن و بولمن ، ۲۰۰۶؛ فریدمن و همکاران ، ۲۰۰۸).

در عمل ممکن است فرض نرمال بودن برقرار نباشد که در این حالت، از تبدیل کردن داده‌ها (با تبدیل‌هایی نظیر تبدیل لگاریتمی) برای بهره‌مندی از ویژگی‌های مدل گرافی گاوسی استفاده می‌شود. بدین منظور، باید همه تبدیل‌ها را بررسی کرد تا تبدیلی که داده‌ها را به نرمال نزدیک می‌کند، حاصل شود. از دیدگاه مدل‌سازی، راه حل ساده‌تر این است که توابع تبدیل، نامعلوم فرض شوند و با روش‌های مناسب برآورد شوند. در این رابطه، فرض نرمال توسط لیو و همکاران (۲۰۰۹) با یک مفصل نیمه‌پارامتری گاوسی [۸] جایگزین شده است. به‌طور مشخص، بردار تصادفی $\boldsymbol{X} = (X_1, \ldots, X_p)$ با بردار تصادفی تبدیل‌شده $\boldsymbol{f}(\boldsymbol{X}) = \big(f_1(X_1), \ldots, f_p(X_p)\big)$ جایگزین شده است که در آن $\boldsymbol{f}(\boldsymbol{X}) \sim N(\boldsymbol{0}, \Sigma)$. این رویکرد به یک تعمیم توزیع نرمال منجر شد که توزیع نرمال ناپارامتری [۹] نامیده شده است. مدل گرافی نرمال ناپارامتری [۱۰] (NGM) انعطاف‌پذیرتر از GGM است و این ویژگی GGM را حفظ می‌کند که روابط استقلال شرطی بین متغیرها در ماتریس دقت $\Omega = \Sigma^{-1}$ در یک فرم گرافی کدگذاری می‌شود. در نتیجه، برآورد NGM به برآورد توابع تبدیل تک متغیره $\{f_j\}_{j=1}^{p}$ و ماتریس دقت Ω بستگی دارد.

در ادبیات مدل‌های گرافی، غالباً فرض بر آن است که داده‌های مشاهده‌شده از یک منبع همگن آمده‌اند و از مدل گرافی گاوسی پارامتری (یوان و لین ، ۲۰۰۷)، مدل گرافی نرمال ناپارامتری (لیو و همکاران ، ۲۰۰۹) یا مدل گرافی نرمال ناپارامتری SKEPTIC (لیو و همکاران ، ۲۰۱۲) پیروی می‌کنند. با این وجود، مشاهده‌ها معمولاً از منبع‌های متفاوت می‌آیند و در طیف گسترده‌ای از کاربردهای دنیای واقعی، ممکن است وابستگی‌های ساختاری

[1] Undirected graphical models
[2] High-dimensional data
[3] Gaussian Graphical Model
[4] Sparse precision matrix
[5] ℓ_1-regularized maximum likelihood
[6] Neighborhood selection
[7] Graphical lasso
[8] Gaussian semiparametric copula
[9] Nonparanormal distribution
[10] Nonparanormal Graphical Model

در کل جامعه وجود داشته باشد. به عبارت دقیق‌تر، ممکن است مشاهده‌ها از خوشه‌ها یا مولفه‌های متفاوت آمده باشند؛ بدون آن‌که اطلاعاتی در مورد عضویت‌شان در خوشه‌ها در دست باشد. بنابراین، کشف زیرجامعه‌های متناهی با وجه اشتراک‌های خاص و برآورد وابستگی‌ها یا استقلال‌های شرطی در این زیرجامعه‌ها، یک چالش اساسی است. خوشه‌بندی مبتنی بر مدل [۱] با استفاده از مدل‌های آمیخته، ابزاری قدرتمند برای مدل‌بندی چنین داده‌هایی فراهم آورده است (لیندسی ، ۱۹۹۵؛ مک لاکلان و پیل ، ۲۰۰۴). ادبیات غنی پیرامون الگوریتم‌های خوشه‌بندی از جمله K –میانگین [۲]، خوشه‌بندی سلسله‌مراتبی [۳] و خوشه‌بندی مبتنی بر مدل وجود دارد (مک کوئین ، ۱۹۶۷).

در روش‌های پیشین، به توزیع آماری داده‌ها توجه نشده است. به‌عنوان مثال، الگوریتم K –میانگین براساس میانگین فاصله اشیاء تا میانگین هر خوشه عمل می‌کند و اشیاء به‌گونه‌ای به خوشه‌ها تخصیص داده می‌شود که میانگین مجموع مربعات فاصله‌ها در خوشه‌ها، کم‌ترین مقدار را داشته باشد. در مقابل، در خوشه‌بندی مبتنی بر مدل، یک مدل آمیخته برای داده‌ها فرض می‌شود و با استفاده از آن، به برآورد برچسب اشیاء که معرف خوشه‌ی آن‌ها است، پرداخته می‌شود. به عبارت دقیق‌تر، خوشه‌بندی مبتنی بر مدل، یک الگوریتم خوشه‌بندی احتمالی است که با آن، میزان شدت باور شخص مبنی بر این که یک نقطه داده شده به یک خوشه خاص تعلق دارد، با تخصیص احتمال‌های پسین به آن نقطه معین ابراز می‌شود.

مدل‌های آمیخته گرافی یک رویکرد خوشه‌بندی مبتنی بر شبکه است که خوشه‌بندی مبتنی بر مدل و مدل‌های گرافی را گرد هم می‌آورد. این مدل‌ها امکان بازسازی خوشه‌ها و برآورد زیرشبکه‌های خوشه‌ای را به‌طور هم‌زمان فراهم آورده است. مدل آمیخته گرافی گاوسی [۴] (GGMM) توسط لاتسی و ویت (۲۰۱۶) معرفی شد که روش‌شناسی و کاربرد مدل GGM را تعمیم داد. برای برآورد نسبت‌های آمیخته و ماتریس‌های دقت مدل GGMM، نسخه جدیدی از الگوریتم امید ریاضی–بیشینه‌سازی [۵] (EM) طراحی شد. اگرچه GGMM نسبت به GGM برای مدل‌بندی داده‌های ناهمگن انعطاف‌پذیرتر و مناسب‌تر است، اما هنوز فرض محدودکننده نرمال را برای مولفه‌های مدل در نظر می‌گیرد. برای تقلیل فرض نرمال بودنِ مولفه‌های مدل، مدل آمیخته گرافی نرمال ناپارامتری [۶] (NGMM) با استفاده از توزیع نرمال ناپارامتری، توسط خلیلی و همکاران (۲۰۲۱) پیشنهاد شد. برآورد ساختار مدل NGMM، به برآورد نسبت‌های آمیخته، ماتریس‌های دقت و تبدیل‌های حاشیه‌ای تک متغیره بستگی دارد.

هدف این مطالعه، بازسازی خوشه‌ها از طریق یک رویکرد خوشه‌بندی مبتنی بر مدل است که مولفه‌های آن توسط مدل گرافی نرمال ناپارامتری تعریف شده است. بدین منظور، یک لگاریتم درستنمایی تاوانیده در نظر گرفته شده است که عبارت تاوان آن به نسبت‌های آمیخته وابسته است. الگوریتم EM مورد استفاده، امکان بازیابی تخصیص خوشه‌ها و برآورد پارامترهای مدل را می‌دهد. در سرتاسر این مطالعه، خوشه‌بندی مبتنی بر مدل توسط

[1] Model-based clustering
[2] K-means
[3] Hierarchical clustering
[4] Gaussian Graphical Mixture Model
[5] Expectation-maximization algorithm
[6] Nonparanormal Graphical Mixture Model

مک لاکلان و پیل (۲۰۰۴) در نظر گرفته شده و فرض بر آن است که داده‌ها از یک مدل متناهی آمیخته نرمال ناپارامتری استخراج شده‌اند که تعداد مولفه‌های آن (K)، معلوم است.

در بخش ۲ پیش زمینه مورد نیاز پیرامون مدل گرافی نرمال ناپارامتری از لیو و همکاران (۲۰۰۹) و مدل آمیخته گرافی نرمال ناپارامتری از خلیلی و همکاران (۲۰۲۱) مرور شده است. سپس، در بخش ۳ رویکرد برآورد درستنمایی ℓ_1-تاوانیده با دو فرم تابع تاوان (متعارف و نامتعارف) از طریق یک الگوریتم EM مبتنی بر مفصل ارائه شده است. بخش ۴ شامل یک مقایسه تجربی بین چندین روش خوشه‌بندی مبتنی بر مدل روی داده‌های شبیه‌سازی‌شده با دو نوع سازوکار آلودگی داده‌ها (قطعی و تصادفی) است. در بخش ۵، روش‌های خوشه‌بندی مورد بحث روی یک مجموعه داده واقعی سرطان سینه مقایسه شده‌اند. بخش پایانی به نتیجه‌گیری و بحث در مورد یافته‌ها می‌پردازد.

۲ پیش زمینه

یک گراف بی‌سو $G = (V, E)$ شامل مجموعه‌ای متناهی از راس‌های $V = \{1, \ldots, p\}$ و مجموعه‌ای از یال‌های $E \subset V \times V$ (که برخی از جفت راس‌های (i, j) را به هم متصل می‌کند) است. مدل‌های گرافی بی‌سو، توزیع توام n مشاهده مستقل از یک بردار تصادفی p-بعدی $\boldsymbol{X} = (X_1, \ldots, X_p)$ متناظر با راس‌های گراف G را به تصویر می‌کشد. زمانی‌که توزیع توام $\boldsymbol{X}$، توزیع گاوسی، نرمال ناپارامتری، آمیخته گاوسی و آمیخته نرمال ناپارامتری فرض شده باشد، ویژگی‌های مدل گرافی مورد نظر برآورد شده است. توزیع نرمال ناپارامتری و رویکرد برآوردیابی جای‌گذاری [1] برای برآورد مدل گرافی نرمال ناپارامتری شرح داده شده است. همچنین به مرور الگوریتم لاسوی گرافی که برای برآورد ماتریس دقت به‌کار می‌رود، پرداخته شده است. در نهایت، مدل آمیخته گرافی نرمال ناپارامتری معرفی شده است.

۲.۱ مدل نرمال ناپارامتری (NPN)

فرض کنید $\{f_j\}_{j=1}^p$ مجموعه‌ای از تابع‌های یکنوا و مشتق‌پذیر تک متغیره نامعلوم است که به‌عنوان تبدیل‌های حاشیه‌ای در نظر گرفته شده‌اند. همچنین فرض کنید $\Sigma \in \mathbb{R}^{p \times p}$ یک ماتریس همبستگی همیشه‌مثبت است که بدون از دست دادن کلیت مساله $diag(\Sigma) = \mathbf{1}$ و $\boldsymbol{\mu} = \mathbf{0}$.

تعریف ۱. (لیو و همکاران ، ۲۰۰۹) بردار تصادفی p-بعدی $\boldsymbol{X} = (X_1, \ldots, X_p)$ دارای توزیع p-بعدی نرمال ناپارامتری[2] است و آن را با نماد $\boldsymbol{X} \sim NPN(\boldsymbol{f}, \Sigma)$ نمایش می‌دهند، هرگاه مجموعه‌ای از تبدیلات یکنوا و مشتق‌پذیر تک متغیره $\{f_j\}_{j=1}^p$ موجود باشد، به‌نحوی‌که

$$\boldsymbol{f}(\boldsymbol{X}) = \big(f_1(X_1), \ldots, f_p(X_p)\big) := (Z_1, \ldots, Z_p) = \boldsymbol{Z} \sim N(\mathbf{0}, \Sigma). \qquad (۱)$$

[1] Plug-in estimation approach
[2] Nonparanormal

نرمال بودن حاشیه‌ای همیشه با تبدیلات به‌دست می‌آید، بنابراین مدل (۱) فرض می‌کند که توزیع توامِ آن متغیرهای تبدیل‌شده، نرمال است. از آن‌جایی‌که توزیع‌های حاشیه‌ای تک متغیره نرمال هستند، مولفه‌های ماتریس همبستگی Σ، دقیقاً همبستگی زوجی بین متغیرها را نشان می‌دهند. اگر f_jها توابع مشتق‌پذیر باشند، تابع چگالی احتمال توام $\boldsymbol{X}$ به فرم

$$P_X(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\big(\boldsymbol{f}(\boldsymbol{x})\big)^T \Sigma^{-1}\big(\boldsymbol{f}(\boldsymbol{x})\big)\right\} \prod_{j=1}^{p} |f'_j(x_j)|. \qquad (2)$$

خواهد بود که شناساپذیر نیست. برای این‌که این خانواده شناساپذیر باشد، فرض شده است که

$$\mu_j = E(Z_j) = E(X_j) = 0, \qquad \sigma_j^2 = Var(Z_j) = Var(X_j) = 1, \qquad j = 1, \ldots, p.$$

با فرض این‌که $\boldsymbol{Z} = (Z_1, \ldots, Z_p) = \big(f_1(X_1), \ldots, f_p(X_p)\big)$ از یک توزیع نرمال پیروی می‌کند، واضح است که $\omega_{ij} = 0$ اگر و تنها اگر $Z_i \perp Z_j | \boldsymbol{Z}_{\setminus\{i,j\}}$، که ω_{ij} نماد درایه (i,j) از ماتریس دقت $\Omega = \Sigma^{-1}$ است. این ایده توسط لیو و همکاران (۲۰۰۹) به صورت

$$\omega_{ij} = 0 \iff Z_i \perp Z_j | \boldsymbol{Z}_{\setminus\{i,j\}} \iff X_i \perp X_j | \boldsymbol{X}_{\setminus\{i,j\}} \iff (i,j) \notin E, \qquad (3)$$

کامل شد، که در آن $\boldsymbol{X}_{\setminus\{i,j\}} := \{X_k : k \neq i, j\}$. از این‌رو، برای برآورد گراف نرمال ناپارمتری، کافی است تابع‌های تک متغیره نامعلوم $\{f_j\}_{j=1}^p$ برآورد شوند و سپس، مکان درایه‌های غیرصفر ماتریس Ω مشخص شوند. در نتیجه، روند طبیعی برآورد مدل گرافی نرمال ناپارامتری، یک رویکرد برآوردیابی جای‌گذاری است که شامل دو گام زیر است:

۱- برای برآورد تبدیل f_j روی متغیر X_j، از یک برآوردگر طبیعی به‌صورت $\tilde{f}_j$ استفاده می‌شود.

۲- از یک تعمیم روش برآوردیابی مدل گرافی گاوسی (نظیر لاسوی گرافی) روی داده‌های تبدیل‌شده، برای برآورد ماتریس دقت استفاده می‌شود.

حال فرض کنید $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$ یک نمونه تصادفی از $NPN(\boldsymbol{0}, \Sigma)$ باشد که $\boldsymbol{X}^i = (X_1^i, \ldots, X_p^i)$. همچنین $\tilde{\boldsymbol{f}}(\boldsymbol{X}^i) = (\tilde{f}_1(X_1^i), \ldots, \tilde{f}_p(X_p^i))$ نمونه تبدیل‌شده باشد که $\tilde{\boldsymbol{f}}(\boldsymbol{X}^1), \ldots, \tilde{\boldsymbol{f}}(\boldsymbol{X}^n)$. ماتریس همبستگی نمونه‌ای R برای متغیرهای تصادفی تبدیل‌شده، به‌عنوان برآورد ماتریس همبستگی نامعلوم Σ، به‌صورت

$$\hat{\Sigma} = R = \sum_{i=1}^{n} \big(\tilde{\boldsymbol{f}}(\boldsymbol{X}^i)\big)\big(\tilde{\boldsymbol{f}}(\boldsymbol{X}^i)\big)^T,$$

تعریف می‌شود، که در آن مولفه‌های ماتریس R برابرند با $r_{jk} = \frac{\sum_{i=1}^{n} \tilde{f}_j(x_j^i)\tilde{f}_k(x_k^i)}{\sqrt{\sum_{i=1}^{n} \tilde{f}_j^2(x_j^i)}\sqrt{\sum_{i=1}^{n} \tilde{f}_k^2(x_k^i)}}$.

برآوردگر ماکسمیم درستنمایی ماتریس دقت با معکوس کردن ماتریس همبستگی به‌دست می‌آید؛ یعنی $\hat{\Omega} = R^{-1}$. Σ^{-1} زمانی‌که تعداد متغیرها بسیار بیشتر از تعداد مشاهدات باشد ($n \ll p$)، R یک ماتریس تکین است و نمی‌توان از آن برای برآورد Ω استفاده کرد. در این حالت، برآوردهای تُنُک را می‌توان با اِعمال یک تاوان ℓ_1 روی مولفه‌های Ω به‌دست آورد. لگاریتم درستنمایی تاوانیده به صورت

$$\ell_p(\Omega) = \log|\Omega| - tr(R\Omega) - \lambda\|\Omega\|_1, \qquad (۴)$$

است، که در آن λ یک پارامتر تنظیم‌کننده [1] برای کنترل سطح تُنُکی است و $\|\Omega\|_1 = \sum_{i\neq j} \omega_{ij}$. یکی از روش‌های یافتن برآورد ماکسیمم درستنمایی تاوانیده [2] (PMLE) برای ماتریس دقت Ω در (۴)، استفاده از نسخه گرافی الگوریتم لاسو [3](glasso) است که توسط فریدمن و همکاران (۲۰۰۸) به فرم

$$\hat{\Omega}^{\lambda}_{glasso} = \underset{\Omega>0}{\arg\max} \Big\{ \log|\Omega| - tr\big(R\Omega\big) - \lambda\|\Omega\|_1 \Big\}, \qquad (۵)$$

پیشنهاد شد، که در آن $\Omega > 0$ بیان‌گر این است که Ω یک ماتریس همیشه‌مثبت است.

۲.۲ مدل آمیخته گرافی نرمال ناپارامتری (NGMM)

فرض کنید $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$ یک نمونه داده شده به اندازه n از K مولفه آمیخته باشد؛ به نحوی که $\boldsymbol{X}^i = (X_1^i, \ldots, X_p^i) \in \mathbb{R}^p$، $i = 1, \ldots, n$. همچنین فرض کنید که هر $\boldsymbol{X}^i$ از یکی از K مولفه با انتخاب تصادفی Y_i از مجموعه $\{1, \ldots, K\}$ با احتمال π_k استخراج شده باشد. π_k نماد نسبت‌های آمیخته است که $0 < \pi_k < 1$ و مجموع آن‌ها در شرط $\sum_k \pi_k = 1$ صدق می‌کند. بنابراین Y_i یک متغیر پنهان چندجمله‌ای با پارامترهای π_k است؛ به‌طوری‌که $P(Y_i = k) = \pi_k$ برای $k = 1, \ldots, K$. هدف، مدل‌بندی داده‌های داده شده با تعیین توزیع توام $P\big(\boldsymbol{X}^i, Y_i\big) = P\big(\boldsymbol{X}^i \big| Y_i\big)P(Y_i)$ است. هر خوشه، به‌طور جداگانه با در نظر گرفتن یک توزیع نرمال ناپارامتری مدل‌بندی می‌شود، به‌طوری‌که $\boldsymbol{X}^i \big| Y_i = k \sim NPN(\boldsymbol{f}_k, \Sigma_k)$ تابع چگالی $\boldsymbol{X}^i$ در خوشه kام باشد.

تعریف ۲. (خلیلی و همکاران ، ۲۰۲۱) بردار تصادفی p-بعدی $\boldsymbol{X}^i = (X_1^i, \ldots, X_p^i)$ در مدل آمیخته نرمال ناپارامتری صدق می‌کند و آن را با نماد $\boldsymbol{X}^i \sim MNPN\{(\pi_k, \boldsymbol{f}_k, \Sigma_k)\}_{k=1}^K$ نمایش می‌دهند، هرگاه تابع

[1] Regularization tuning parameter
[2] Penalized Maximum likelihood estimate
[3] Graphical lasso

چگالی توام $\boldsymbol{X}^i = (X_1^i, \ldots, X_p^i)$ به فرم

$$f(\boldsymbol{x}^i \big| \Psi) = \sum_{k=1}^{K} \pi_k NPN(\boldsymbol{x}^i \big| \xi_k) = \sum_{k=1}^{K} \pi_k NPN(\boldsymbol{x}^i \big| \boldsymbol{f}_k, \Sigma_k), \qquad (۶)$$

باشد، که در آن $NPN(.\,|\,\xi_k)$ نماد چگالی نرمال ناپارامتری به فرم (۲) است، $\Psi = (\Pi, \Xi)$ بردار همه پارامترهای نامعلوم است و $\xi_k = (\boldsymbol{f}_k, \Sigma_k)$ پارامترهای مولفه kام مدل برای $k = 1, \ldots, K$ است.

در مدل (۶)، فضای پارامتر را می‌توان به فرم

$$\Psi = \Pi \times \Xi, \qquad (۷)$$

نوشت، که در آن $\Pi = \left\{ \{\pi_k\}_{k=1}^{K} \Big| \, 0 < \pi_k < 1, \sum_k \pi_k = 1, k = 1, \ldots, K \right\}$ و
$\Xi = \left\{ \{f_k\}_{k=1}^{K}, \{\Omega_k\}_{k=1}^{K} \Big| \, \Omega_k > 0, k = 1, \ldots, K \right\}$.

مدل آمیخته نرمال ناپارامتری که در تعریف ۲ ارائه شده است، در صورتی شناساپذیر است که برای هر دو پارامتر $\psi, \psi^* \in \Psi$ و برای همه مقدارهای ممکن $\boldsymbol{x}$، رابطه زیر برقرار باشد

$$\sum_{k=1}^{K} \pi_k NPN(\boldsymbol{x}|\,\xi_k) = \sum_{k=1}^{K^*} \pi_k^* NPN(\boldsymbol{x}|\,\xi_k^*) \Longleftrightarrow$$
$$K = K^*, \quad \pi_k = \pi_k^*, \quad NPN(\boldsymbol{x}|\,\xi_k) = NPN(\boldsymbol{x}|\,\xi_k^*).$$

تساوی $NPN(\boldsymbol{x}|\,\xi_k) = NPN(\boldsymbol{x}|\,\xi_k^*)$ زمانی برای همه مقدارهای $\boldsymbol{x}$ برقرار است که $f_{j(k)}$ها میانگین و واریانس را حفظ کنند؛ یعنی

$$\mu_{j(k)} = E(Z_{j(k)}) = E(X_j) = 0 \quad , \quad \sigma_{j(k)}^2 = Var(Z_{j(k)}) = Var(X_j) = 1 \qquad (۸)$$

به‌ازای هر $j \in \{1, \ldots, p\}$ و $k \in \{1, \ldots, K\}$.

۳ برآورد و خوشه‌بندی همزمان برای مدل آمیخته گرافی

در این بخش، ابتدا بردارهای تبدیل برآورد شده‌اند. سپس برآورد PMLE و به‌طور همزمان برآورد خوشه‌ها با استفاده از یک الگوریتم EM بر مبنای مفصل ارائه شده است.

۳.۱ برآورد بردارهای تبدیل حاشیه‌ای

فرض کنید $\boldsymbol{X}^1,\ldots,\boldsymbol{X}^n$ نمونه‌ای به اندازه n از آمیخته متناهی از مدل نرمال ناپارامتری باشد؛ یعنی $\boldsymbol{X}^i=(X_1^i,\ldots,X_p^i)\overset{i.i.d.}{\sim} MNPN\{(\pi_k,\boldsymbol{f}_k,\Sigma_k)\}_{k=1}^K$. همچنین فرض کنید $F_j(x)$ نماد تابع توزیع حاشیه‌ای X_j، $j=1,\ldots,p$، باشد. بردار تبدیل متناظر با مولفه kام را به‌صورت

$$\boldsymbol{f}_k(\boldsymbol{X})=\big(f_{1(k)}(X_1),\ldots,f_{p(k)}(X_p)\big):=(Z_{1(k)},\ldots,Z_{p(k)})=\boldsymbol{Z}_k\sim N(\boldsymbol{0},\Sigma_k)$$

در نظر بگیرید که مولفه‌های قطری ماتریس همبستگی برابر با یک هستند؛ $\sigma_{jj(k)}=1$، $j=1,\ldots,p$ و $k=1,\ldots,K$. آن‌گاه

$$F_j(x)=\sum_{k=1}^K\pi_kP(X_j\le x)=\sum_{k=1}^K\pi_kP\big(f_{j(k)}(X_j)\le f_{j(k)}(x)\big)=\sum_{k=1}^K\pi_k\Phi\big(f_{j(k)}(x)\big),$$

که در آن Φ نماد تابع توزیع تجمعی نرمال استاندارد تک متغیره است. تحت شرایط شناساپذیری در (۸) و با فرض این‌که بردارهای تبدیلات برای همه مولفه‌ها یکسان‌اند ($f_{j(k)}=f_{j(k')}$ ، $\forall k,k'\in\{1,\ldots,K\}$) که می‌توان از اندیس k صرف‌نظر کرد)، تبدیل حاشیه‌ای متغیر jام در همه مولفه‌ها برابر $f_j(x)=\Phi^{-1}\big(F_j(x)\big)$ است و برآوردگر تبدیل حاشیه‌ای f_j به صورت

$$\tilde{f}_j(x)=\Phi^{-1}\big(\tilde{F}_j(x)\big),\qquad(۹)$$

تعریف می‌شود، که در آن $\tilde{F}_j$ یک برآوردگر F_j است. مشابه لیو و همکاران (۲۰۰۹)، برای اجتناب از بزرگیِ واریانس برآوردگرها در ابعاد بالا، از برآوردگر $\tilde{F}_j(x)=n\hat{F}_j(x)\big/(n+1)$ استفاده می‌کنیم، که در آن $\hat{F}_j(t)=\frac{1}{n}\sum_{i=1}^n I(X_j^i\le t)$، تابع توزیع تجربی حاشیه‌ای X_j است و $I(\cdot)$ تابع نشانگر و t یک مقدار ثابت است.

۳.۲ درستنمایی تاوانیده مبتنی بر مدل

فرض کنید $\boldsymbol{X}^1,\ldots,\boldsymbol{X}^n$ نمونه‌ای تصادفی از $MNPN\{(\pi_k,\boldsymbol{f}_k,\Sigma_k)\}_{k=1}^K$ باشد و همچنین فرض کنید $\tilde{\boldsymbol{f}}(\boldsymbol{X}^1),\ldots,\tilde{\boldsymbol{f}}(\boldsymbol{X}^n)$، داده‌های تبدیل شده متناظر با این نمونه با استفاده از (۹) باشد. طبق رویکرد برآوردیابی جای‌گذاری، برای پیاده‌سازی PMLE، مشاهدات با داده‌های تبدیل شده جای‌گزین می‌شوند. بنابراین مساله برآورد مدل NGMM، به مساله برآورد ماتریس‌های دقت $\Omega_k=\Sigma_k^{-1}$ و نسبت‌های آمیخته π_k برای $k=1,\ldots,K$

تقلیل می‌یابد و به دنبال آن، فضای پارامتری (۷) با فضای پارامتری $\Theta = \Pi \times \Omega$ جای‌گزین می‌شود، که در آن

$$\Pi = \{\{\pi_k\}_{k=1}^{K} | 0 < \pi_k < 1, \sum_k \pi_k = 1, k = 1, \ldots, K\}$$
$$\Omega = \{\{\Omega_k\}_{k=1}^{K}\} | \Omega_k > 0, k = 1, \ldots, K\}.$$

حال می‌توان تابع درستنمایی داده‌های ناتمام را به‌صورت

$$L(\Theta) = \prod_{i=1}^{n} \Big(\sum_{k=1}^{K} \pi_k NPN\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i) \Big| \Omega_k^{-1}\big) \Big),$$

نوشت، که در آن $\Theta = \{(\pi_k, \Omega_k) : k = 1, \ldots, K\}$ مجموعه تمام پارامترهای نامعلوم است.

تابع لگاریتم درستنمایی برابر است با

$$\ell(\Theta) = \sum_{i=1}^{n} \log \Big(\sum_{k=1}^{K} \pi_k NPN\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i) \Big| \Omega_k^{-1}\big) \Big). \qquad (۱۰)$$

برای افزایش سطح تُنُکی، یک تکنیک ماکسیمم درستنمایی تاوانیده با اعمال یک تاوان ℓ_1 روی هریک از K ماتریس دقت توسط ژو و همکاران (۲۰۰۹) و موخرجی و هیل (۲۰۱۱)، برای خوشه‌بندی با استفاده از مدل گرافی گاوسی ارائه شده است. همچنین یک تابع تاوان مشابه توسط خلیلی و چن (۲۰۰۷) و استدلر و همکاران (۲۰۱۰) برای آمیخته‌ی متناهی از مدل‌های رگرسیونی تاوانیده ارائه شده است. با اِعمال این تاوان بر لگاریتم درستنمایی (۱۰)، لگاریتم درستنمایی تاوانیده به‌صورت

$$\ell_p(\Theta) = \sum_{i=1}^{n} \log \Big(\sum_{k=1}^{K} \pi_k NPN\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i) \Big| \Omega_k^{-1}\big) \Big) - \frac{n}{2} \lambda \sum_{k=1}^{K} \pi_k^{\gamma} \|\Omega_k\|_1, \qquad (۱۱)$$

تعریف می‌شود، که در آن λ یک پارامتر تنظیم‌کننده است، γ یک پارامتر دودویی با مقادیر ۰ و ۱ است که فرم عبارت تاوان را کنترل می‌کند، $\|.\|_1$ نرم ℓ_1 است و $\|\Omega_k\|_1 = \sum_{i \neq j} |\omega_{ij(k)}|$. وقتی $\gamma = 0$ باشد، عبارت تاوان به نسبت‌های آمیخته π_k وابسته نیست و بنابراین، لگاریتم درستنمایی (۱۱)، به فرم متعارف لگاریتم درستنمایی تاوانیده که توسط لاتسی و ویت (۲۰۱۶) برای مدل آمیخته گرافی گاوسی و همچنین توسط خلیلی و همکاران (۲۰۲۱) برای مدل آمیخته گرافی نرمال ناپارامتری مورد استفاده قرار گرفته، تبدیل می‌شود. اما وقتی $\gamma = 1$ باشد، لگاریتم درستنمایی (۱۱)، به عبارت تاوان از هر K خوشه به نسبت π_k وزن می‌دهد. بنابراین، تاوان ℓ_1 (۱۱) فرم کلی‌تر دارد؛ زیرا می‌تواند به نسبت‌های آمیخته بستگی داشته باشد یا نداشته باشد. در این مطالعه، این دو فرم تابع تاوان،

برای خوشه‌بندی و برآورد پارامترهای مدل آمیخته گرافی گاوسی و مدل آمیخته گرافی نرمال ناپارامتری به‌کار رفته است.

۳.۳ الگوریتم EM مبتنی بر مفصل

برای داده‌های مستقل $\boldsymbol{C} = \{(\boldsymbol{x}^i, y_i), i = ۱, \ldots, n\}$، لگاریتم درستنمایی تاوانیده به‌صورت

$$\ell_{p,c}(\Theta) = \sum_{i=۱}^{n}\sum_{k=۱}^{K} 1_{\{Y_i=k\}} \log\Big(\pi_k NPN\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\Big|\,\Omega_k^{-۱}\big)\Big) - \frac{n}{۲}\lambda\sum_{k=۱}^{K}\pi_k^{\gamma}\|\Omega_k\|_{۱},$$

است، که در آن اگر مشاهده $\boldsymbol{x}^i$ به خوشه k متعلق باشد، $Y_i = k$ است. بنابراین، $1_{\{Y_i=k\}}$ نشان می‌دهد که اگر مشاهده $\boldsymbol{x}^i$ به خوشه k متعلق باشد، از Θ_k استفاده می‌شود.

گام E. در گام E از الگوریتم EM، احتمال‌های شرطی این‌که $\boldsymbol{X}^i$ از خوشه k آمده باشد، به شرط آن‌که برآوردهای جاری پارامترهای $\Theta^{(t)}$ و داده‌های تبدیل شده، داده شده باشند، محاسبه می‌شود. ابتدا کمیت

$$\begin{aligned}
Q(\Theta|\,\Theta^{(t)}) &= E_{Y_i}\Big[\ell_{p,c}(\Theta)|\,\{\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\}_{i=۱}^{n}, \Theta^{(t)}\Big] \\
&= \sum_{i=۱}^{n}\sum_{k=۱}^{K} E_{Y_i}\Big[1_{\{Y_i=k\}}\Big|\,\tilde{\boldsymbol{f}}(\boldsymbol{x}^i), \Theta^{(t)}\Big] \\
&\times \Big[\log\pi_k + \log\big(NPN(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\Big|\,\Omega_k^{-۱})\big)\Big] - \frac{n}{۲}\lambda\sum_{k=۱}^{K}\pi_k^{\gamma}\|\Omega_k\|_{۱} \\
&= \sum_{i=۱}^{n}\sum_{k=۱}^{K}\tau_k^{i(t)}\Big[\log\pi_k + \log\big(NPN(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\Big|\,\Omega_k^{-۱})\big)\Big] - \frac{n}{۲}\lambda\sum_{k=۱}^{K}\pi_k^{\gamma}\|\Omega_k\|_{۱},
\end{aligned} \tag{۱۲}$$

محاسبه می‌شود، که در آن $\tau_k^{i(t)}$ احتمال پسین تعلق مشاهده $\boldsymbol{x}^i$ به خوشه k است. طبق قضیه بیز

$$\tau_k^{i(t)} = P\big(\,Y_i = k|\,\tilde{\boldsymbol{f}}(\boldsymbol{x}^i), \Theta^{(t)}\big) = \frac{\pi_k^{(t)} NPN\big(\,\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\Big|\,\Omega_k^{-۱(t)}\big)}{\sum_{j=۱}^{K}\pi_j^{(t)} NPN\big(\,\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\Big|\,\Omega_j^{-۱(t)}\big)}. \tag{۱۳}$$

برای اجرای خوشه‌بندی مبتنی بر مدل، هر مشاهده $\boldsymbol{x}^i$ را به خوشه k تخصیص می‌دهند، هرگاه $\tau_k^{i(t)} > \tau_{k'}^{i(t)}$، $\forall k' \neq k$ و $k' = ۱, \ldots, K$.

گام M. برای به‌دست آوردن برآوردهای به‌روز رسانی‌شده $\Theta^{(t+۱)}$، کمیت $Q(\,\Theta|\,\Theta^{(t)})$ نسبت به : $\Theta = \{(\pi_k, \Omega_k)$

$k = ۱, \ldots, K\}$ ماکسیمم می‌شود. بنابراین، گام **M** به دو مساله بهینه‌سازی تقسیم می‌شود؛ ماکسیمم‌سازی نسبت به π_k و ماکسیمم‌سازی نسبت به Ω_k.

۱- گام $\boldsymbol{M}$ برای π_k: وقتی $\gamma = ۰$، نسبت‌های آمیخته π_k در عبارت تاوان ظاهر نمی‌شوند، لذا قید $\pi_K = ۱ - \sum_{k=۱}^{K-۱} \pi_k$ و $\pi_k > ۰$ به‌کار برده می‌شود. در این حالت، با ماکسیمم‌سازی (۱۲) نسبت به π_k، به‌روز رسانی استاندار EM مشابه خلیلی و همکاران (۲۰۲۱) به‌دست می‌آید

$$\pi_k^{(t+۱)} = \frac{\sum_{i=۱}^{n} \tau_k^{i(t)}}{n}. \tag{۱۴}$$

به ازای $\gamma = ۱$، نسبت‌های آمیخته π_k در عبارت تاوان ظاهر می‌شوند. در این حالت، از به‌روز رسانی استاندارد (۱۴) استفاده می‌کنیم. اگر این به‌روز رسانی استاندارد، $Q(\Theta|\,\Theta^{(t)})$ را بهبود بخشد، آنگاه برای به‌دست آوردن ماکسیمم (۱۱) کافی است. اگرچه بهبود در این مورد تضمین نشده است، اما این روش در کاربرد به‌خوبی عمل می‌کند، هم‌چنان که در خلیلی و چن (۲۰۰۷) نشان داده شده است.

۲- گام $\boldsymbol{M}$ برای Ω_k: برای ماکسیمم کردن (۱۲) نسبت به ماتریس دقت Ω_k، عباراتی که به Ω_k بستگی ندارند، حذف می‌شوند. بنابراین، کمیت $Q(\Theta|\,\Theta^{(t)})$ به کمیت $Q(\Omega_k)$ تقلیل می‌یابد

$$\begin{aligned} Q(\Omega_k) &= \sum_{i=۱}^{n} \tau_k^{i(t)} \log\left(NPN(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\Big|\,\Omega_k^{-۱})\right) - \frac{n}{۲}\lambda\pi_k^{\gamma}\|\Omega_k\|_{۱} \\ &= \sum_{i=۱}^{n} \tau_k^{i(t)} \Big[\frac{۱}{۲}\log|\Omega_k| - \frac{۱}{۲}\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\big)^T\Omega_k\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\big)\Big] - \frac{n}{۲}\lambda\pi_k^{\gamma}\|\Omega_k\|_{۱} \\ &= \frac{۱}{۲}\sum_{i=۱}^{n} \tau_k^{i(t)} \Big[\log|\Omega_k| - tr\big(R_k\Omega_k\big) - \tilde{\lambda}_k\|\Omega_k\|_{۱}\Big]. \end{aligned}$$

بنابراین، به‌روز رسانی Ω_k به‌صورت

$$\Omega_k^{(t+۱)} = \underset{\Omega_k > ۰}{\arg\max}\Big\{\log|\Omega_k| - tr\big(R_k^{(t)}\Omega_k\big) - \tilde{\lambda}_k^{(t)}\|\Omega_k\|_{۱}\Big\}, \tag{۱۵}$$

به‌دست می‌آید، که در آن $R_k^{(t)}$ به‌روز رسانی ماتریس همبستگی تجربی وزنی است که برابر است با

$$\begin{aligned} R_k^{(t)} &= \frac{\sum_{i=۱}^{n} \tau_k^{i(t)}\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\big)\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\big)^T}{\sum_{i=۱}^{n} \tau_k^{i(t)}}, \\ \tilde{\lambda}_k^{(t)} &= n\lambda\frac{\big(\pi_k^{(t+۱)}\big)^{\gamma}}{\sum_{i=۱}^{n} \tau_k^{i(t)}}. \end{aligned} \tag{۱۶}$$

از (۱۴) و (۱۶) نتیجه می‌شود

$$\tilde{\lambda}_k^{(t)} = \begin{cases} \frac{\lambda}{\pi_k^{(t+1)}} & \gamma = 0 \\ \lambda & \gamma = 1 \end{cases} \tag{۱۷}$$

رابطه (۱۵) به فرم لاسوی گرافی در (۵) است که در آن، ماتریس همبستگی نمونه R با یک ماتریس همبستگی تجربی وزنی R_k و پارامتر تنظیم‌کننده λ با یک پارامتر تنظیم‌کننده مقیاس $\tilde{\lambda}_k$ جای‌گزین شده است. بنابراین، می‌توان از الگوریتم لاسوی گرافی برای حل این مساله بهینه‌سازی استفاده کرد. توجه داشته باشید که وقتی $\gamma = 0$، فرم متعارف لگاریتم درستنمایی تاوانیده (۱۱) از یک پارامتر تنظیم‌کننده λ استفاده می‌کند، درحالی‌که به‌روز رسانی EM (۱۵) یک پارامتر تنظیم‌کننده متناسب با عکس نسبت آمیخته خوشه را ارائه می‌دهد. برای $\gamma = 1$، عکس این مطلب برقرار است.

۳.۴ انتخاب پارامتر تنظیم‌کننده

میزان سطح تنکی در ماتریس دقت و مدل گرافی مربوطه، توسط پارامتر تنظیم‌کننده λ کنترل می‌شود. روش‌های مختلف انتخاب پارامتر تنظیم‌کننده همراه با فرم‌های متفاوت تاوان ℓ_1، منجر به رویکردهای برآوردیابی متفاوت می‌شود. فرم کلی معیار اطلاع بیزی [1] (BIC) به‌صورت

$$BIC(\lambda) = -2\ell(\hat{\Theta}^\lambda) + \log(n) df^\lambda, \tag{۱۸}$$

است، که در آن $\ell(\cdot)$ لگاریتم درستنمایی غیرتاوانیده (۱۰)، $\hat{\Theta}^\lambda$ برآورد درستنمایی تاوانیده با پارامتر تنظیم‌کننده λ و df درجه آزادی برای اندازه‌گیری پیچیدگی مدل است. برای تعیین درجه آزادی در مدل گرافی گاوسی، کمیت

$$df^\lambda = K(p+1) - 1 + \sum_{k=1}^{K} \#\left\{(i,j) : i \leq j, \hat{\omega}_{ij(k)}^\lambda \neq 0\right\}, \tag{۱۹}$$

توسط یوان و لین (۲۰۰۷) پیشنهاد شده است، که در آن $\hat{\omega}_{ij(k)}^\lambda$ درایه (i,j) در $\hat{\Omega}_k^\lambda$ است. λ با مینیمم‌سازی نمره‌های BIC موجود برای هر λ، انتخاب می‌شود؛ یعنی $\hat{\lambda} = \arg\min_\lambda BIC(\lambda)$. توجه کنید که مقدار بهینه‌ی پارامتر تنظیم‌کننده کمتر تحت تاثیر تخصیص خوشه‌ها است، ولی به شدت به ویژگی‌های کلی داده‌ها نظیر p، n و K بستگی دارد.

[1] Bayesian Information Criterion

۳.۵ الگوریتم برآورد تخصیص خوشه‌ها و پارامترهای مدل

الگوریتم ۱. تخصیص خوشه‌ها و برآورد پارامترها

ورودی: n مشاهده مستقل $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$ که $\boldsymbol{X}^i = (X_1^i, \ldots, X_p^i)$، $i = 1, \ldots, n$.

گام ۱- جای‌گزینی مشاهده‌های هر متغیر X_j، $j = 1, \ldots, p$، با مقدار استاندارد شده آن‌ها.

گام ۲- برآورد تبدیل‌های f_j از رابطه $\tilde{f}_j(x) = \Phi^{-1}\big(\frac{1}{n+1}\sum_{i=1}^{n} I(X_j^i \le x)\big)$، $j = 1, \ldots, p$.

گام ۳- مقدار دهی اولیه $\Theta^{(0)} = \big\{(\pi_k^{(0)}, \Omega_k^{(0)})\big\}_{k=1}^{K}$:

۱.۳- تخصیص تصادفی هر مشاهده $\boldsymbol{x}^i$ به یکی از K خوشه منوط به اندازه معین $n_k^{(0)} = n/2$.

۲.۳- تعیین $\pi_k^{(0)} = n_k^{(0)}/n = 0/5$ برای $k = 1, \ldots, K$ که $n_k^{(0)}$ تعداد مشاهده‌های تخصیص داده شده اولیه به خوشه k است.

۳.۳- تعیین $\Omega_k^{(0)}$، $k = 1, \ldots, K$، برای ماتریس دقت خوشه k با استفاده از لاسوی گرافی

$$\Omega_k^{(0)} = \underset{\Omega_k > 0}{\arg\max} \Big\{ \log|\Omega_k| - tr\big(R_k^{(0)}\Omega_k\big) - \lambda\|\Omega_k\|_1 \Big\},$$

که $R_k^{(0)} = \sum_{i=1}^{n} \big(\tilde{\boldsymbol{f}}_k(\boldsymbol{X}^i)\big)\big(\tilde{\boldsymbol{f}}_k(\boldsymbol{X}^i)\big)^T$، ماتریس همبستگی نمونه اولیه در خوشه k است.

گام ۴- گام E : محاسبه احتمال‌های پسین $\tau_k^{i(t)} = \frac{\pi_k^{(t)} NPN\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\big|\Omega_k^{-1(t)}\big)}{\sum_{j=1}^{K} \pi_j^{(t)} NPN\big(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i)\big|\Omega_j^{-1(t)}\big)}$.

گام ۵- گام $\boldsymbol{M}$: محاسبه برآوردهای به‌روز رسانی‌شده پارامترها $\Theta^{(t+1)} = \big\{(\pi_k^{(t+1)}, \Omega_k^{(t+1)})\big\}_{k=1}^{K}$:

۱.۵- محاسبه $\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_k^{i(t)}}{n}$ برای $k = 1, \ldots, K$.

۲.۵- تعیین به‌روز رسانی $\boldsymbol{EM}$ برای ماتریس دقت خوشه k، $k = 1, \ldots, K$، از رابطه

$$\Omega_k^{(t+1)} = \underset{\Omega_k > 0}{\arg\max} \Big\{ \log|\Omega_k| - tr\big(R_k^{(t)}\Omega_k\big) - \tilde{\lambda}_k^{(t)}\|\Omega_k\|_1 \Big\},$$

که $R_k^{(t)} = \frac{\sum_{i=1}^{n} \tau_k^{i(t)} (\tilde{\boldsymbol{f}}(\boldsymbol{x}^i))(\tilde{\boldsymbol{f}}(\boldsymbol{x}^i))^T}{\sum_{i=1}^{n} \tau_k^{i(t)}}$ و $\tilde{\lambda}_k^{(t)} = \frac{\lambda}{\pi_k^{(t+1)}} 1_{\{\gamma=0\}} + \lambda 1_{\{\gamma=1\}}$.

گام ۶- تخصیص هر مشاهده $\boldsymbol{x}^i$ به خوشه k با بزرگترین احتمال $\tau_k^{i(t)}$، $i = 1, \ldots, n$.

گام ۷- تکرار گام‌های ۴ تا ۶ یا توقف آن‌ها در صورت برآورده شدن یکی از شرایط زیر:

(i) ماکسیمم تعداد تکرارها *(T)* حاصل شود: $t > T$.

(ii) برای خوشه k، مینیمم اندازه خوشه *(n_{min})* حاصل شود: $\sum_{i=1}^{n} \tau_k^{i(t)} < n_{\min}$.

(iii) تغییر نسبی در لگاریتم درستنمایی، ناچیز باشد: $\epsilon \le \Big|\ell_p(\Theta)^{(t)}/\ell_p(\Theta)^{(t-1)} - 1\Big|$. **خروجی.** برآورد تخصیص خوشه‌ها، نسبت‌های آمیخته و ماتریس‌های دقت خوشه‌ای.

۴ مطالعه شبیه‌سازی

نتایج تجربی خوشه‌بندی مبتنی بر مدل آمیخته گرافی نرمال ناپارامتری و خوشه‌بندی مبتنی بر مدل آمیخته گرافی گاوسی، برای دو فرم کلی تابع تاوان ℓ_1 (متعارف و نامتعارف)، مقایسه شده است. بدین منظور، ابتدا فرایند تولید داده‌ها و روند مقایسه کمی شرح داده شده است. سپس مقدار پارامتر تنظیم کننده انتخاب شده برای هر روش، نمایش داده شده است. در ادامه، میزان تطبیق خوشه‌های بازسازی‌شده در هر روش با خوشه‌های اصلی و همچنین میزان تطبیق پارامترهای برآورد شده از خوشه‌های بازسازی‌شده در هر روش با پارامترهای واقعی از خوشه‌های اصلی، مقایسه شده است.

۴.۱ تولید داده

برای تولید داده از مدل آمیخته نرمال ناپارامتری، داده‌های p-بعدی شامل دو خوشه ($K = ۲$) با دو ماتریس دقت تنک معلوم در نظر گرفته شده است. بدین منظور، ابتدا دو گراف تنک تولید می‌شود و سپس، ماتریس‌های دقت Ω_1 و Ω_2 متناظر با این دو گراف به‌دست می‌آیند. در ادامه، n نمونه مستقل از $N(\mathbf{0}, \Sigma_k)$ که $\Sigma_k = \Omega_k^{-1}$ و $k = ۱, ۲$، استخراج می‌شود. بعد از آن، با استفاده از تبدیل‌های $\{f_{j(k)}\}_{j=1}^{p}$، داده‌های نرمال $N(\mathbf{0}, \Sigma_k)$ به داده‌های نرمال ناپارامتری $NPN(\boldsymbol{f}_k, \Sigma_k)$ تبدیل می‌شوند. در نهایت، مقادیر متغیر برنولی $Y_i \sim Ber(۱, ۰/۵)$، $i = ۱, \ldots, n$، تولید و براساس مقادیر آن، داده‌های $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$ از مدل آمیخته نرمال ناپارامتری دو مولفه‌ای با احتمال‌های برابر $\pi_k = ۰/۵$، تولید می‌شوند. جزئیات الگوهای تولید گراف در ادامه شرح داده شده است.

الگوی I (بسته نرم‌افزاری Huge[1]). الگوی گراف بدین ترتیب تولید می‌شود که احتمال آن‌که بین هر زوج راس‌های (i, j) $(i \neq j)$، یالی وجود داشته باشد، برابر است با

$$P\big((i,j) \in E\big) = \sqrt{۰/۵\delta_p} I(\delta_p < ۰/۵) + \big(۱ - \sqrt{۰/۵(۱ - \delta_p)}\big) I(\delta_p \geq ۰/۵),$$

که $\delta_p = \min(۱, ۳/p)$ و p تعداد راس‌ها است. مجموعه یال E شامل هر زوج راس (i, j) می‌شود، اگر و تنها اگر، مقدار احتمال $P\big((i,j) \in E\big)$، از احتمال متناظر با یک توزیع یکنواخت دو متغیره در مربع $[۰, ۱]^2$ کمتر باشد. معکوس ماتریس همبستگی $\Omega_1 = \big(\omega_{ij(1)}\big)_{p \times p}$ برای خوشه ۱ به‌صورت زیر ساخته می‌شود

$$\omega_{ij(1)} = \begin{cases} |e| + ۰/۱ + u & i = j \\ ۰/۳ & (i,j) \in E \\ ۰ & \text{سایر نقاط} \end{cases}$$

[1] http://cran.r-project.org/web/packages/huge.

که در آن، u عددی مثبت است که برای کنترل بزرگی همبستگی‌های جزئی، به عناصر قطری ماتریس دقت اضافه شده است (مقدار پیش‌فرض ۰/۱ است). برای تضمین همیشه‌مثبت بودن معکوس ماتریس همبستگی، عناصر قطری $\Omega_{۱}$، با مقدار قدر مطلق کوچکترین مقدار ویژه $\Omega_{۱}$ (که با نماد $|e|$ نشان داده شده است)، جایگزین شده‌اند.

الگوی II (مینشاوسن و بولمن ، ۲۰۰۶). الگوی گراف این‌گونه ایجاد می‌شود که مکان هر گره به‌طور یکسان و یکنواخت در یک مربع دوبعدی $[۰,۱]^{۲}$ توزیع شده است. به‌عبارت دقیق‌تر، هر اندیس $j \in \{۱,\dots,p\}$ با یک نقطه داده دومتغیره $\left(U_j^{(۱)}, U_j^{(۲)}\right) \in [۰,۱]^{۲}$ در تناظر است، به‌طوری‌که $U_{۱}^{(m)},\dots,U_p^{(m)} \sim uniform[۰,۱]$ برای $m = ۱,۲$. هر جفت از راس‌های (i,j) ابتدا با احتمال

$$P\big((i,j)\in E\big) = \frac{۱}{\sqrt{۲\pi}}\exp\Big(-\frac{\|u_i-u_j\|^{۲}}{۲s}\Big),$$

در مجموعه راس E قرار می‌گیرد، که در آن $u_i =: \left(u_i^{(۱)}, u_i^{(۲)}\right)$ مقدار مشاهده شده $\left(U_i^{(۱)}, U_i^{(۲)}\right)$ و $\|\cdot\|$ نماد فاصله اقلیدسی بین زوج متغیرها است. $s = ۰/۱۲۵$ پارامتری است که سطح تنکی گراف تولید شده را کنترل می‌کند. برای دستیابی به سطح تنکی مطلوب، ماکسیمم درجه گراف، یعنی حداکثر تعداد یال‌های متصل به هر گره، به ۴ محدود شده است و یال‌های اضافی متصل به هر گره، به‌طور تصادفی حذف می‌شوند. بنابراین، معکوس ماتریس همبستگی $\Omega_{۲} = \left(\omega_{ij(۲)}\right)_{p\times p}$ برای خوشه ۲ به‌صورت

$$\omega_{ij(۲)} = \begin{cases} ۱ & i=j \\ ۰/۲۴۵ & (i,j)\in E \\ ۰ & \text{سایر نقاط} \end{cases}$$

تشکیل می‌شود که در آن همبستگی جزئی بین راس‌های مجاور، ۰/۲۴۵ در نظر گرفته شده است. مقادیر مطلق کمتر از ۰/۲۵، همیشه‌مثبت بودن معکوس ماتریس همبستگی را تضمین می‌کند. برای به‌دست آوردن ماتریس همبستگی، ماتریس کواریانس $\Sigma = \Omega^{-۱}$ حاصل، به ماتریس همبستگی تبدیل می‌شود؛ به‌نحوی‌که همه عناصر قطری آن برابر با ۱ شود.

داده‌های نرمال و داده‌های نرمال ناپارامتری از مدل‌های زیر تولید می‌شود:

- مدل (۱)- مدل آمیخته گاوسی چند متغیره با دو مولفه با استفاده از تبدیل خطی روی داده‌ها.
- مدل (۲)- مدل آمیخته نرمال ناپارامتری چند متغیره با دو مولفه با استفاده از تبدیل‌های توانی تک متغیره یکسان روی هر بعد، ارائه شده توسط لیو و همکاران (۲۰۰۹).
- مدل (۳)- مدل آمیخته نرمال ناپارامتری چند متغیره با دو مولفه با استفاده از تبدیل‌های تابع توزیع (CDF) تک متغیره یکسان روی هر بعد، ارائه شده توسط لیو و همکاران (۲۰۰۹).

در مدل‌های (۲) و (۳)، برای هر مولفه $k = ۱, ۲$، ابتدا n داده مستقل از $N(\mathbf{0}, \Sigma_k)$ که $\Sigma_k = \Omega_k^{-۱}$، تولید می‌شود و سپس، با استفاده از بردارهای تبدیل $\boldsymbol{f}_k^{-۱} = \left(f_{۱(k)}^{-۱}, \ldots, f_{p(k)}^{-۱}\right) =: \boldsymbol{g}_k$، داده‌ها به $NPN(\boldsymbol{f}_k, \Sigma_k)$ تبدیل می‌شود. همان‌طور که در بخش ۳.۱ فرض شد، بردارهای تبدیل روی هر مولفه، تابع‌های یکسان هستند؛ یعنی $\boldsymbol{f}_۱ = \boldsymbol{f}_۲$ یا $f_{j(۱)} = f_{j(۲)}$. بنابراین از اندیس (k) برای تابع‌های تبدیل، صرف‌نظر شده است. همچنین در مدل‌های (۲) و (۳) تابع تبدیل‌های تک متغیره یکسان روی هر متغیر X_j، برای $j = ۱, \ldots, p$، در نظر گرفته شده است؛ یعنی $f_۱ = f_۲ = \cdots = f_p = f$. در مدل‌های (۲) و (۳)، دو نسخه مختلف از $g_j := f_j^{-۱}$ به‌کار برده شده است که این دو تبدیل، در تعریف ۹ و تعریف ۱۰ از مقاله لیو و همکاران (۲۰۰۹) تعریف شده‌اند.

همچنین دو نوع سازوکار آلودگی داده‌ها در سطح $r \in (۰, ۱)$ در نظر گرفته شد؛ قطعی و تصادفی. برای آلودگی قطعی (سطری)، تعداد $[nr]$ مشاهده با بردار قطعی $(+۵, -۵, +۵, -۵, \ldots) \in \mathbb{R}^p$ جای‌گزین می‌شود، به‌نحوی‌که اعداد $+۵$ و -۵ به‌صورت متناوب تکرار شوند. برای آلودگی تصادفی (ستونی)، $[nr]$ مولفه از هر بعد براساس یک توزیع یکنواخت، با اعداد $+۵$ یا -۵ به‌صورت تصادفی جای‌گزین می‌شوند. این دو نوع آلودگی، برای تجزیه و تحلیل داده‌های علمی مدرن، واقع‌بینانه‌تر هستند.

برای تولید داده، دو بُعد $p = ۱۵, ۲۰$ و دو اندازه نمونه $n = ۳۰, ۱۵۰$ در نظر گرفته شده که با آلودگی قطعی، ۵ درصد داده دورافتاده $(r = ۰/۰۵)$ و با آلودگی تصادفی، ۲۰ درصد داده دورافتاده $(r = ۰/۲۰)$ جای‌گزین شده است. همچنین، معیار توقف یکسان $T = ۱۰۰$، $n_{min} = ۴$ و $\epsilon = ۱۰^{-۴}$، برای الگوریتم EM اِعمال شده است. همه محاسبات با نرم‌افزار R انجام شده است.

به‌طور کلی، روش‌های زیر با هم مقایسه شده‌اند:

(i) $GGMM_۰$: خوشه‌بندی مبتنی بر مدل GGMM با استفاده از فرم متعارف لگاریتم درستنمایی $\ell_۱$-تاوانیده از لاتسی و ویت (۲۰۱۶).

(i) $GGMM_۱$: خوشه‌بندی مبتنی بر مدل GGMM با استفاده از فرم نامتعارف لگاریتم درستنمایی $\ell_۱$-تاوانیده از موخرجی و هیل (۲۰۱۱).

(iii) $NGMM_۰$: خوشه‌بندی مبتنی بر مدل NGMM با استفاده از فرم متعارف لگاریتم درستنمایی $\ell_۱$-تاوانیده از خلیلی و همکاران (۲۰۲۱).

(iv) $NGMM_۱$: خوشه‌بندی مبتنی بر مدل NGMM با استفاده از فرم نامتعارف لگاریتم درستنمایی $\ell_۱$-تاوانیده پیشنهاد شده در این مقاله.

شایان ذکر است لاتسی و ویت (۲۰۱۶) و خلیلی و همکاران (۲۰۲۱) از فرم متعارف لگاریتم درستنمایی $\ell_۱$-تاوانیده تنها برای برآورد پارامترهای GGMM و NGMM استفاده کردند.

۴.۲ مقادیر پارامتر تنظیم‌کننده

بعد از تولید داده از مدل‌های (۱) تا (۳)، گام‌های زیر برای انتخاب پارامتر تنظیم‌کننده طی شده‌اند:
۱- تولید $K = ۲$ خوشه‌نما با تخصیص تصادفی مشاهدات به هر خوشه.
۲-محاسبه مقادیر اولیه پارامتر برای هر دو خوشه‌نما:
- $\hat{\pi}_k^{(۰)}$ به‌عنوان نسبت اندازه‌های نمونه در خوشه‌نمای k تعیین می‌شود.
- برای ۱۰ مقدار متفاوت λ، $\hat{\Omega}_k^{(۰)\lambda}$ به‌طور همزمان با استفاده از الگوریتم لاسوی گرافی (۵) در خوشه‌نمای k به‌دست می‌آید. این‌کار با استفاده از تابع huge.glasso در بسته نرم‌افزاری Huge انجام شده است.
۳- با استفاده از این برآوردها، نمره‌های BIC محاسبه می‌شوند و برای انتخاب یک پارامتر λ مشترک بین هر دو خوشه، مینیمم می‌شوند.
۴- گام‌های ۱ تا ۳، ۵۰ بار تکرار می‌شوند و از مقدارهای λ به‌دست آمده، میانگین گرفته می‌شود تا یک مقدار نهایی حاصل شود.

میانگین این مقدارها برای هر روش و به ازای دو مقدار متفاوت برای n و p، در جداول ۱ و ۲ منعکس شده است. طبق (۲۰)، برای روش‌های $GGMM_۱$ و $NGMM_۱$ ($\gamma = ۱$)، به‌روز رسانی‌های EM برای پارامترهای تنظیم‌کننده $\tilde{\lambda}_k$ برابر با λ است که در جداول ۱ و ۲ نیز این برابری برقرار است. همان‌طور که از فرم BIC در (۱۸) و (۱۹) انتظار می‌رفت، مقدارهای λ به دست آمده با افزایش n، کاهش می‌یابند. همچنین با افزایش سطح آلودگی داده‌ها (r)، انحراف استاندارد مقدارهای λ افزایش می‌یابد، اما میانگین مقدارهای λ در برابر افزایش آلودگی، تقریباً استوار است.

۴.۳ انتساب خوشه

روش‌های خوشه‌بندی دارای دو بخش معیار و تکنیک است. معیار، به هر خوشه یک مقدار عددی اختصاص می‌دهد که گویای مطلوبیت نسبی روش مورد نظر است. شاخص رَند [۱](RI)، یک معیار عینی برای سنجش میزان تشابه بین تخصیص‌های خوشه‌های واقعی و خوشه‌های برآورد شده است که در بازه $(۰, ۱)$ مقدار می‌پذیرد. $RI = ۰$ بدین معنی است که الگوریتم خوشه‌بندی مورد نظر در برآورد ساختار جامعه شکست خورده است و در مقابل، $RI = ۱$ به معنی انطباق کامل خوشه‌های برآورد شده با خوشه‌های واقعی است (رند ، ۱۹۷۱). شکل‌های ۱ و ۲ شاخص‌های رند به‌دست آمده در هر روش از خوشه‌بندی ۵۰ مجموعه داده شبیه‌سازی شده با سه تبدیل داده (خطی، توانی و تابع توزیع) و دو سازوکار آلودگی (قطعی با ۵ درصد آلودگی و تصادفی با ۲۰ درصد آلودگی)، برای دو مقدار متفاوت n و p را نشان می‌دهد. تفسیر نتایج بدین شرح است:
الف- داده‌های نرمال و بدون داده دورافتاده ($r = ۰$). زمانی‌که تبدیل خطی روی داده‌های نرمال اعمال می‌شود، روش‌های مورد بررسی در برآورد خوشه‌های واقعی تفاوت معنی‌داری ندارند و هر چهار روش عملکرد مشابهی دارند.

[1] Rand index

جدول ۱. متوسط مقدارهای پارامتر تنظیم‌کننده برای هر روش روی ۵۰ مجموعه داده شبیه‌سازی‌شده در بُعد $p = ۱۵$ و اندازه نمونه $n = ۳۰$. انحراف استانداردها داخل پرانتزها نوشته شده‌اند.

		$GGMM_۰$			$GGMM_۱$	$NGMM_۰$			$NGMM_۱$
تبدیل	r	λ	$\tilde{\lambda}_۱$	$\tilde{\lambda}_۲$	$\lambda = \tilde{\lambda}_k$	λ	$\tilde{\lambda}_۱$	$\tilde{\lambda}_۲$	$\lambda = \tilde{\lambda}_k$
	۰٫۰۰	۰٫۶۴	۱٫۲۶	۱٫۳	۰٫۶۵	۰٫۶۳	۱٫۲۳	۱٫۳۰	۰٫۶۳
		(۰٫۰۴)	(۰٫۱۳)	(۰٫۰۸)	(۰٫۰۴)	(۰٫۰۵)	(۰٫۱۶)	(۰٫۱۰)	(۰٫۰۵)
خطی	۰٫۰۵	۰٫۷۰	۱٫۳۷	۱٫۷۱	۰٫۷۲	۰٫۶۶	۱٫۳۱	۱٫۳۱	۰٫۶۴
		(۰٫۰۹)	(۰٫۴۸)	(۰٫۳۹)	(۰٫۰۸)	(۰٫۰۵)	(۰٫۱۱)	(۰٫۱۰)	(۰٫۰۵)
	۰٫۲۰	۰٫۶۲	۱٫۴۵	۱٫۳۲	۰٫۶۲	۰٫۶۰	۱٫۱۰	۱٫۷۴	۰٫۶۰
		(۰٫۶۳)	(۰٫۶۵)	(۰٫۴۳)	(۰٫۷۳)	(۰٫۵۳)	(۰٫۷۴)	(۰٫۶۷)	(۰٫۶۷)
	۰٫۰۰	۰٫۷۶	۱٫۵۱	۱٫۵۶	۰٫۷۶	۰٫۶۳	۱٫۲۲	۱٫۳۰	۰٫۶۲
		(۰٫۰۶)	(۰٫۱۸)	(۰٫۰۹)	(۰٫۰۵)	(۰٫۰۵)	(۰٫۱۷)	(۰٫۱۰)	(۰٫۰۶)
توانی	۰٫۰۵	۰٫۸۲	۱٫۶۵	۱٫۶۹	۰٫۸۰	۰٫۶۵	۱٫۳۱	۱٫۳۰	۰٫۶۴
		(۰٫۰۶)	(۰٫۳۲)	(۰٫۲۱)	(۰٫۰۶)	(۰٫۰۵)	(۰٫۱۱)	(۰٫۱۰)	(۰٫۰۵)
	۰٫۲۰	۰٫۶۳	۱٫۳۸	۱٫۲۸	۰٫۸۲	۰٫۵۹	۱٫۳۴	۱٫۳۱	۰٫۶۰
		(۰٫۸۵)	(۰٫۳۵)	(۰٫۳۵)	(۰٫۸۲)	(۰٫۵۸)	(۰٫۳۱)	(۰٫۲۸)	(۰٫۶۵)
	۰٫۰۰	۰٫۶۱	۱٫۲۱	۱٫۲۵	۰٫۶۵	۰٫۶۴	۱٫۲۵	۱٫۳۰	۰٫۶۳
		(۰٫۰۷)	(۰٫۱۶)	(۰٫۱۲)	(۰٫۰۶)	(۰٫۰۵)	(۰٫۱۵)	(۰٫۰۹)	(۰٫۰۵)
تابع توزیع	۰٫۰۵	۰٫۷۰	۱٫۳۵	۱٫۷۲	۰٫۷۲	۰٫۶۶	۱٫۳۲	۱٫۳۱	۰٫۶۴
		(۰٫۱۰)	(۰٫۴۸)	(۰٫۳۸)	(۰٫۰۹)	(۰٫۰۶)	(۰٫۱۲)	(۰٫۱۱)	(۰٫۰۵)
	۰٫۲۰	۰٫۶۲	۱٫۳۸	۱٫۳۲	۰٫۶۱	۰٫۶۰	۱٫۲۰	۱٫۱۴	۰٫۶۰
		(۰٫۶۷)	(۰٫۵۷)	(۰٫۴۰)	(۰٫۷۳)	(۰٫۵۲)	(۰٫۴۹)	(۰٫۴۵)	(۰٫۶۷)

ب- داده‌های غیرنرمال و بدون داده دورافتاده ($r = ۰$). زمانی‌که با تبدیل توانی توزیع داده‌ها با توزیع نرمال اختلاف پیدا می‌کند، عملکرد روش $GGMM_۰$ افت می‌کند، ولی همچنان سه روش دیگر عملکرد مشابه دارند. در مورد تبدیل تابع توزیع، هر چهار روش عملکرد نزدیک به هم در خوشه‌بندی داده‌ها دارند.

ج- داده‌های نرمال با داده دورافتاده ($r \in \{۰/۰۵, ۰/۲۰\}$). زمانی‌که با سازوکار آلودگی قطعی، ۵ درصد از داده‌های نرمال یا با سازوکار آلودگی تصادفی، ۲۰ درصد از داده‌ها با داده‌های دورافتاده جایگزین می‌شوند، عملکرد روش متعارف $GGMM_۰$ (و در یک مورد، روش $GGMM_۱$) افت می‌کند، ولی روش‌های نرمال ناپارامتری $NGMM_۰$ و $NGMM_۱$، عملکرد بهتری دارند.

د- داده‌های غیرنرمال با داده دورافتاده ($r \in \{۰/۰۵, ۰/۲۰\}$). در این حالت‌ها، روش‌های مورد بررسی بهتر مقایسه می‌شوند. زمانی‌که اندازه نمونه کوچک است، روش‌های نامتعارف $GGMM_۱$ و $NGMM_۱$ عملکرد بهتری نسبت به روش‌های متعارف $GGMM_۰$ و $NGMM_۰$ دارند و در برخی موارد، $NGMM_۱$ بهتر از $GGMM_۱$ است. با افزایش اندازه نمونه، روش‌های مبتنی بر نرمال ناپارامتری $NGMM_۰$ و $NGMM_۱$، نسبت به روش‌های مبتنی بر نرمال $GGMM_۰$ و $GGMM_۱$، در مقابل آلودگی داده‌ها استوارتر هستند و عملکرد بهتری دارند.

جدول ۲. متوسط مقدارهای پارامتر تنظیم‌کننده برای هر روش روی ۵۰ مجموعه داده شبیه‌سازی‌شده در بُعد $p = ۲۰$ و اندازه نمونه $n = ۱۵۰$. انحراف استانداردها داخل پرانتزها نوشته شده‌اند.

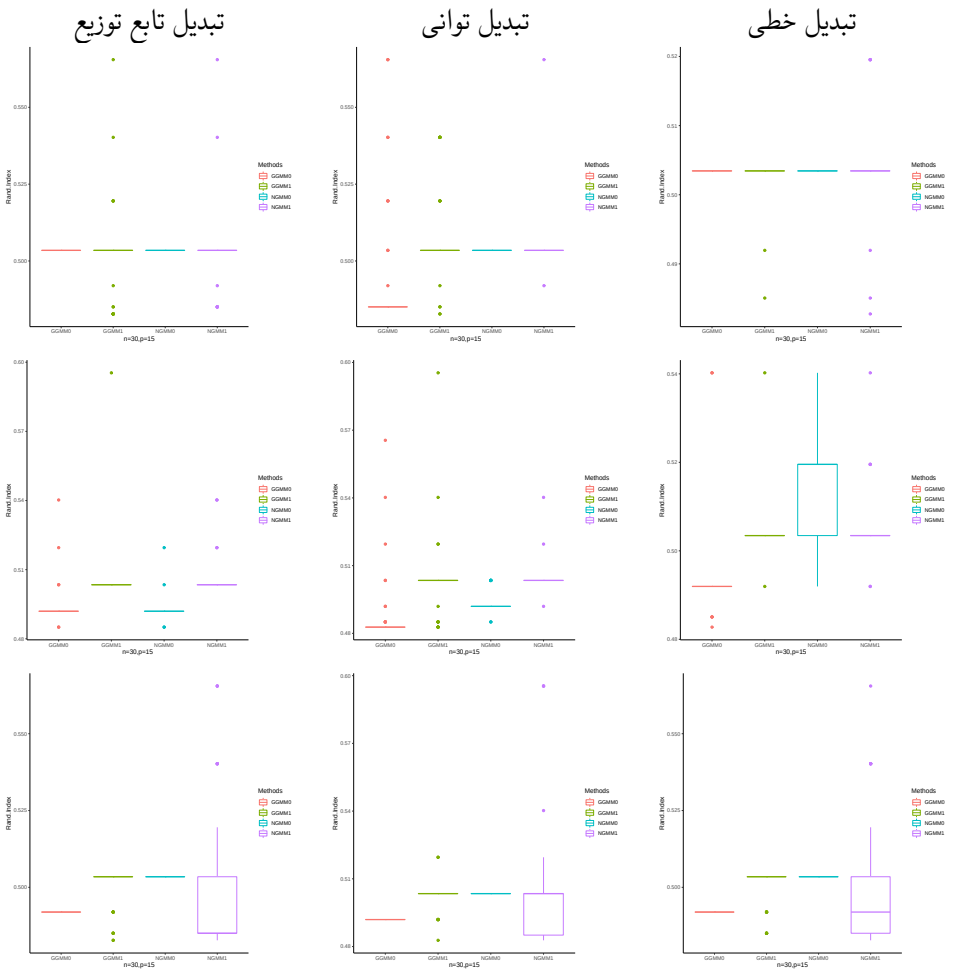
تبدیل	r	$GGMM_۰$			$GGMM_۱$	$NGMM_۰$			$NGMM_۱$
		λ	$\tilde{\lambda}_۱$	$\tilde{\lambda}_۲$	$\lambda = \tilde{\lambda}_k$	λ	$\tilde{\lambda}_۱$	$\tilde{\lambda}_۲$	$\lambda = \tilde{\lambda}_k$
	۰٫۰۰	۰٫۳۸	۰٫۷۶	۰٫۷۶	۰٫۳۷	۰٫۳۵	۰٫۷۱	۰٫۷۱	۰٫۳۶
		(۰٫۰۶)	(۰٫۱۲)	(۰٫۱۱)	(۰٫۰۷)	(۰٫۰۷)	(۰٫۱۵)	(۰٫۱۵)	(۰٫۰۷)
خطی	۰٫۰۵	۰٫۶۴	۷٫۸۶	۶٫۶۲	۰٫۶۴	۰٫۵۳	۱٫۴۰	۱٫۳۶	۰٫۵۱
		(۰٫۰۵)	(۷٫۰۲)	(۶٫۲۵)	(۰٫۰۶)	(۰٫۰۵)	(۰٫۶۷)	(۰٫۶۸)	(۰٫۰۵)
	۰٫۲۰	۰٫۳۱	۱٫۴۷	۲٫۱۲	۰٫۳۰	۰٫۳۰	۳٫۷۷	۳٫۲۲	۰٫۲۹
		(۰٫۴۴)	(۱٫۴۸)	(۱٫۵۱)	(۰٫۴۰)	(۰٫۳۷)	(۱٫۵۵)	(۲٫۴)	(۰٫۳۰)
	۰٫۰۰	۰٫۳۴	۰٫۹۷	۱٫۶۳	۰٫۳۴	۰٫۳۳	۰٫۶۶	۰٫۶۷	۰٫۳۴
		(۰٫۰۹)	(۰٫۹۲)	(۲٫۴۵)	(۰٫۱۱)	(۰٫۰۸)	(۰٫۱۷)	(۰٫۱۵)	(۰٫۰۷)
توانی	۰٫۰۵	۰٫۷۲	۵٫۷۳	۵٫۱۵	۰٫۷۳	۰٫۵۳	۱٫۴۹	۱٫۱۲	۰٫۵۱
		(۰٫۰۶)	(۵٫۸۱)	(۴٫۵۲)	(۰٫۰۷)	(۰٫۰۵)	(۰٫۵۶)	(۰٫۵۵)	(۰٫۰۵)
	۰٫۲۰	۰٫۳۱	۱٫۴۸	۱٫۵۸	۰٫۲۹	۰٫۳۰	۶٫۸۷	۲٫۵۱	۰٫۳۰
		(۰٫۰۴)	(۱٫۲۲)	(۱٫۲۳)	(۰٫۰۳)	(۰٫۰۴)	(۳٫۴۵)	(۲٫۶۳)	(۰٫۰۴)
	۰٫۰۰	۰٫۳۶	۰٫۷۲	۰٫۷۲	۰٫۳۵	۰٫۳۴	۰٫۶۷	۰٫۶۸	۰٫۳۵
		(۰٫۰۵)	(۰٫۱۱)	(۰٫۱۱)	(۰٫۰۵)	(۰٫۰۸)	(۰٫۱۶)	(۰٫۱۵)	(۰٫۰۷)
تابع	۰٫۰۵	۰٫۶۳	۷٫۷۸	۵٫۹۷	۰٫۶۳	۰٫۵۳	۱٫۳۴	۱٫۳۴	۰٫۵۱
توزیع		(۰٫۰۵)	(۶٫۹۰)	(۵٫۹۹)	(۰٫۰۶)	(۰٫۰۵)	(۰٫۶۲)	(۰٫۶۲)	(۰٫۰۵)
	۰٫۲۰	۰٫۳۱	۱٫۷۶	۱٫۸۲	۴۰٫۳۰	۰٫۳۰	۲٫۵۰	۱٫۰۸	۰٫۲۹
		(۰٫۰۵)	(۱٫۵۵)	(۱٫۵۳)	(۰٫۰۴)	(۰٫۰۴)	(۱٫۷۷)	(۱٫۹۹)	(۰٫۰۳)

۴.۴ مقایسه برآورد ماتریس‌های دقت

در این بخش، دقت روش‌ها در برآورد درایه‌های ماتریس‌های دقت خوشه‌ها مقایسه شده است. دقت با استفاده از نُرم $\left\|\cdot\right\|_۱$ محاسبه می‌شود که در آن $\sum_{k=۱}^{K} \left\|\hat{\Omega}_k - \Omega_k\right\|_۱$، ماتریس دقت خوشه برآورد شده k و Ω_k، ماتریس دقت خوشه واقعی k است. نتایج در جدول ۳ آورده شده است. عملکرد ضعیف روش متعارف $GGMM_۰$ در خوشه‌بندی داده‌ها موجب شد که این روش، در برآورد ماتریس‌های دقت خوشه‌ها نیز ضعیف‌ترین عملکرد را داشته باشد. در مقابل، روش نامتعارف $NGMM_۱$ بالاترین دقت را نشان داده است. به‌طور کلی، روش‌های نامتعارف $GGMM_۱$ و $NGMM_۱$، دقت بالاتری نسبت به روش‌های متعارف $GGMM_۰$ و $NGMM_۰$ در برآورد پارامترهای مدل‌های آمیخته مورد بررسی داشته‌اند.

۵ کاربرد داده واقعی

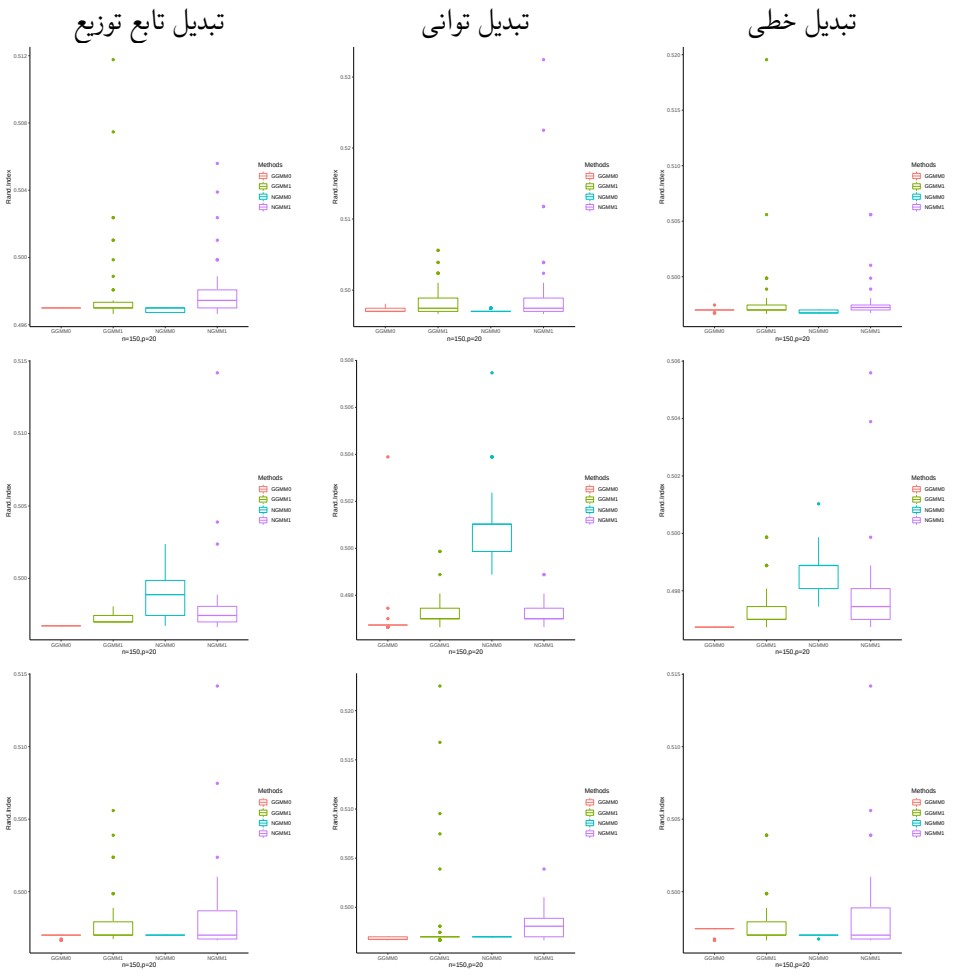
سرطان سینه، شایع‌ترین سرطان و دومین عامل مرگ و میر ناشی از سرطان در بین زنان است. از این‌رو، نظارت بر شیوع سرطان سینه برای دولت‌ها و سازمان‌های بهداشتی، از اهمیت بالایی برخوردار است. در تحقیقات علمی، داده‌های سرطان سینه به عنوان داده‌های ناهمگن شناخته شده است. روش‌های خوشه‌بندی روی داده‌های تشخیص



شکل ۱. مقایسه نتایج انتساب خوشه با استفاده از شاخص رند، روی ۵۰ مجموعه داده شبیه‌سازی‌شده با اندازه نمونه $n = ۳۰$ و بُعد $p = ۱۵$ برای روش‌های (به ترتیب از چپ به راست) $GGMM_۰$، $GGMM_۱$، $NGMM_۰$ و $NGMM_۱$. داده‌ها فاقد آلودگی (سطر اول)، داده‌ها با سازوکار آلودگی قطعی ۵ درصد (سطر دوم) و داده‌ها با سازوکار آلودگی تصادفی ۲۰ درصد (سطر سوم).

سرطان سینه ویسکانسین[1] (WDBC) (منگاسریان و همکاران ، ۱۹۹۵) اعمال شده‌اند.[2] این مجموعه داده شامل

[1] Wisconsin Diagnostic Breast Cancer
[2] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).



شکل ۲. مقایسه نتایج انتساب خوشه با استفاده از شاخص رند، روی ۵۰ مجموعه داده شبیه‌سازی‌شده با اندازه نمونه $n = ۱۵۰$ و بُعد $p = ۲۰$ برای روش‌های (به ترتیب از چپ به راست) $GGMM_۰$، $GGMM_۱$، $NGMM_۰$ و $NGMM_۱$. داده‌ها فاقد آلودگی (سطر اول)، داده‌ها با سازوکار آلودگی قطعی ۵ درصد (سطر دوم) و داده‌ها با سازوکار آلودگی تصادفی ۲۰ درصد (سطر سوم).

هسته‌های سلولی به‌دست آمده از تصویرهای دیجیتالی با سوزن ظریف آسپیراسیون[۱] (FNA) از هر توده سرطان سینه است. داده‌ها در ۱۰ ویژگی روی ۵۶۹ بیمار اندازه‌گیری شده که برای هر ویژگی، ۳ عدد به عنوان میانگین، خطای استاندارد و میانگین سه مقدار بزرگ‌تر ثبت شده است. بنابراین، این مجموعه داده شامل ۵۶۹ سطر، ۳۰ ستون اصلی

[1] Fine Needle Aspirate

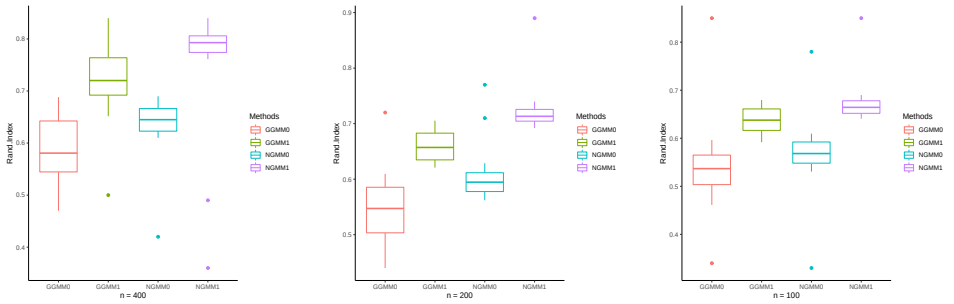
جدول ۳. مقایسه دقت برآورد ماتریس‌های دقت هر دو خوشه برای هر روش روی ۵۰ مجموعه داده شبیه‌سازی‌شده. انحراف استانداردها داخل پرانتزها نوشته شده‌اند.

$NGMM_۱$	$NGMM_۰$	$GGMM_۱$	$GGMM_۰$	r	تبدیل	(p,n)
۸٫۷۸(۰٫۱۱)	۹٫۷۹(۰٫۱۰)	۸٫۸۳(۰٫۰۹)	۹٫۸۰(۰٫۰۸)	۰٫۰۰		
۸٫۷۹(۰٫۱۲)	۹٫۸۴(۰٫۱۲)	۸٫۹۴(۰٫۱۸)	۱۰٫۰۱(۰٫۱۳)	۰٫۰۵	خطی	
۸٫۷۲(۰٫۱۴)	۱۰٫۵۸(۰٫۲۹)	۸٫۷۵(۰٫۱۶)	۹٫۸۳(۰٫۱۳)	۰٫۲۰		
۸٫۷۶(۰٫۱۵)	۹٫۷۹(۰٫۱۱)	۹٫۰۳(۰٫۱۱)	۱۰٫۰۸(۰٫۱۱)	۰٫۰۰		
۸٫۸۰(۰٫۱۲)	۹٫۸۴(۰٫۱۲)	۹٫۱۰(۰٫۱۵)	۱۰٫۱۹(۰٫۱۰)	۰٫۰۵	توانی	(۱۵, ۳۰)
۸٫۷۱(۰٫۱۴)	۹٫۸۲(۰٫۱۷)	۸٫۷۵(۰٫۱۷)	۱۰٫۶۱(۰٫۲۰)	۰٫۲۰		
۸٫۷۸(۰٫۱۰)	۹٫۸۰(۰٫۱۰)	۸٫۸۲(۰٫۱۳)	۹٫۷۴(۰٫۱۷)	۰٫۰۰		
۸٫۸۰(۰٫۱۱)	۹٫۸۴(۰٫۱۳)	۸٫۹۳(۰٫۲۱)	۱۰٫۰(۰٫۱۶)	۰٫۰۵	تابع توزیع	
۸٫۷۲(۰٫۱۴)	۹٫۸۱(۰٫۱۴)	۸٫۷۴(۰٫۱۵)	۱۰٫۵۷(۰٫۲۹)	۰٫۲۰		
۸٫۶۴(۰٫۲۸)	۹٫۵۸(۰٫۳۳)	۸٫۶۴(۰٫۲۸)	۹٫۷۱(۰٫۲۶)	۰٫۰۰		
۹٫۱۰(۰٫۱۵)	۱۰٫۵۱(۰٫۱۱)	۹٫۳۹(۰٫۲۱)	۱۱٫۴۵(۰٫۰۶)	۰٫۰۵	خطی	
۸٫۴۳(۰٫۱۱)	۱۰٫۸۸(۰٫۵۸)	۸٫۴۵(۰٫۱۳)	۱۰٫۳۳(۰٫۲۴)	۰٫۲۰		
۸٫۵۴(۰٫۳۳)	۹٫۴۸(۰٫۳۷)	۸٫۵۱(۰٫۴۲)	۹٫۹۲(۰٫۵۵)	۰٫۰۰		
۹٫۰۹(۰٫۱۵)	۱۰٫۴۵(۰٫۱۰)	۹٫۶۲(۰٫۲۱)	۱۱٫۴۳(۰٫۲۱)	۰٫۰۵	توانی	(۲۰, ۱۵۰)
۸٫۴۴(۰٫۱۴)	۱۰٫۲۲(۰٫۲۰)	۸٫۴۳(۰٫۱۲)	۱۰٫۹۵(۰٫۵۱)	۰٫۲۰		
۸٫۵۸(۰٫۳۳)	۹٫۵۱(۰٫۳۶)	۸٫۶۰(۰٫۱۹)	۹٫۶۲(۰٫۲۵)	۰٫۰۰		
۹٫۱۰(۰٫۱۵)	۱۰٫۴۹(۰٫۱۰)	۹٫۳۵(۰٫۲۳)	۱۱٫۴۰(۰٫۱۶)	۰٫۰۵	تابع توزیع	
۸٫۴۳(۰٫۱۱)	۱۰٫۳۲(۰٫۳۰)	۸٫۴۸(۰٫۱۳)	۱۰٫۹۰(۰٫۵۷)	۰٫۲۰		

و یک ستون «تشخیص» با دو وضعیت تومور بدخیم و خوش‌خیم است.

هدف، پیش‌بینی متغیر «تشخیص» برای ارزیابی بدخیم یا خوش‌خیم بودن تومور هر بیمار است. بدین منظور، ابتدا مشاهدات به‌طور تصادفی به دو خوشه تخصیص داده شد و سپس، روش‌های خوشه‌بندی مبتنی بر مدل $GGMM_۰$، $GGMM_۱$، $NGMM_۰$ و $NGMM_۱$، بر داده‌ها برازش داده شد. برای ارزیابی تاثیر اندازه نمونه بر عملکرد روش‌های خوشه‌بندی مورد مطالعه، ۵۰ نمونه بوت استرپ با اندازه‌های متفاوت $n = ۱۰۰, ۲۰۰, ۴۰۰$، بدون جای‌گذاری از جامعه استخراج شده است.

شکل ۳ نتایج بازیابی خوشه‌ها برای هر روش را منعکس می‌کند. طبق این شکل، روش $NGMM_۱$ نتایج متفاوتی ارائه می‌دهد و به‌طور قابل توجهی بهتر از سایر روش‌های مورد بررسی عمل می‌کند. همچنین روش $GGMM_۱$ اندکی بهتر از روش $NGMM_۰$ عمل می‌کند و روش متعارف $GGMM_۰$ بدترین عملکرد را دارد. شکل ۳ نشان می‌دهد که با افزایش اندازه نمونه، عملکرد روش‌های مورد نظر ارتقا می‌یابد.



$n = ۴۰۰$ $n = ۲۰۰$ $n = ۱۰۰$

شکل ۳. مقایسه نتایج انتساب خوشه با استفاده از شاخص رند، روی ۵۰ نمونه از مجموعه داده سرطان سینه WDBC با اندازه‌های نمونه $n = ۱۰۰, ۲۰۰, ۴۰۰$ و بُعد $p = ۳۰$ برای روش‌های (به ترتیب از چپ به راست) $GGMM_۰$، $GGMM_۱$، $NGMM_۰$ و $NGMM_۱$.

بحث و نتیجه‌گیری

خوشه‌بندی مبتنی بر مدل آمیخته گرافی نرمال ناپارامتری با دو فرم تابع تاوان $\ell_۱$ متعارف و نامتعارف ارائه شد و عملکرد آن با روش خوشه‌بندی مبتنی بر مدل آمیخته گرافی گاوسی با دو فرم تابع تاوان $\ell_۱$ متعارف و نامتعارف، هم از نظر بازسازی خوشه‌ها و هم از نظر برآورد پارامترهای مدل، مقایسه شد. همراه با تاوان $\ell_۱$ استاندارد، یک تاوان جای‌گزین در نظر گرفته شد که به نسبت‌های آمیخته π_k وابسته است، بنابراین، انتخاب تابع تاوان به‌عنوان کلید اصلیِ مقایسه در این مطالعه ظاهر شده است، ولی انتخاب پارامتر تنظیم‌کننده در مقایسه بی‌تاثیر بوده است.

در مطالعه شبیه‌سازی، عملکرد روش‌های مورد بررسی زمانی‌که توزیع داده‌ها نرمال یا غیرنرمال است، بررسی شد. برای مقایسه روش‌ها از نظر استحکام در برابر داده‌های دورافتاده، دو نوع سازوکار آلودگی قطعی و تصادفی در نظر گرفته شد. در برآورد ماتریس‌های دقت خوشه‌ها، روش‌های نامتعارف $GGMM_۱$ و $NGMM_۱$ دقت بالاتری نسبت به روش‌های متعارف $GGMM_۰$ و $NGMM_۰$ داشتند. در بازسازی خوشه‌ها، روش‌های خوشه‌بندی مبتنی بر مدل آمیخته گرافی نرمال ناپارامتری ($NGMM_۰$ و $NGMM_۱$) نسبت به روش‌های مبتنی بر مدل آمیخته گرافی گاوسی، عملکرد بهتری از خود بروز دادند، به خصوص در حضور داده‌های دورافتاده، استوارتر بودند. با این وجود، عملکرد هر چهار روش در بازیابی خوشه‌ها، نزدیک به هم بود و از کارایی مناسب برخوردار نبود ($۰/۴۵ < RI < ۰/۵۵$)، زیرا داده‌ها از مدل‌های آمیخته گاوسی و آمیخته نرمال ناپارامتری با دو مولفه با بردارهای میانگین صفر برای هر مولفه شبیه‌سازی شده‌اند. همچنین الگوهای I و II، ماتریس‌های همبستگی با درایه‌های بین $[۰, ۱]$ تولید می‌کنند. از آن‌جایی‌که داده‌های تولید شده برای هر دو مولفه بسیار نزدیک به هم است، بازیابی خوشه‌ها دشوار است. بنابراین، شاخص‌های رند در محدوده ۰/۵ محاسبه شده‌اند.

در نهایت، روش‌های خوشه‌بندی مورد نظر روی داده‌های سرطان سینه برای تشخیص توده‌های سرطانی خوش‌خیم

و بدخیم بین بیماران، به‌کار برده شد. نتایج حاکی از آن است که روش‌های خوشه‌بندی نامتعارف $GGMM_1$ و $NGMM_1$، به‌طور قابل توجهی عملکرد بهتری در تشخیص وضعیت توده‌های سرطانی داشته‌اند و با افزایش اندازه نمونه، کارایی آن‌ها نیز افزایش می‌یابد ($۰/۷ < RI < ۰/۸$).

به‌طور کلی، نتایج نشان داد که عملکرد روش‌های خوشه‌بندی مبتنی بر مدل، به انتخاب عبارت تاوان و انتخاب مدل بستگی دارد؛ به‌نحوی‌که ترکیبی از مدل آمیخته گرافی نرمال ناپارامتری با عبارت تاوان وابسته به نسبت‌های آمیخته (که در خلال این مطالعه به عنوان روش $NGMM_1$ معرفی شد)، دقیق‌ترین خوشه‌بندی و برآورد پارامترها را ارائه می‌دهد که با افزایش اندازه نمونه، دقت آن بیشتر نمود پیدا می‌کند.

تقدیر و تشکر

نویسندگان از نظرات روشن‌گرانه داوران محترم و پیشنهادهای ارزنده سردبیر محترم، هیئت تحریریه و ویراستار مجله علوم آماری که در بهبود کیفیت مقاله نقش مهمی داشته‌اند، کمال قدردانی و تشکر را می‌نمایند.

مراجع

Friedman, J. H., Hastie, T. and Tibshirani, R. (2008), Sparse Inverse Covariance Estimation with the Graphical Lasso, *Biostatistics*, **9**, 432-441.

Khalili, A. and Chen, J. (2007), Variable Selection in Finite Mixture of Regression Models, *Journal of the American Statistical Association*, **102**, 1025-1038.

Khalili, A., Eskandari, F. and Nematollahi, N. (2021), Estimation of Undirected Graph with Finite Mixture of Nonparanormal Distribution, *Journal of Statistical Theory and Practice*, **15**, 1-18.

Lindsay, B. G. (1995), Mixture Models: Theory, Geometry and Applications, *In NSF-CBMS Regional Conference Series in Probability and Statistics*, 1-163.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012), High-dimensional Semiparametric Gaussian Copula Graphical Models, *The Annals of Statistics*, **40**, 2293-2326.

Liu, H., Lafferty, J. and Wasserman, L. (2009), The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs, *Journal of Machine Learning Research*, **10**, 2295-2328.

Lotsi, A. and Wit, E. (2016), Sparse Gaussian Graphical Mixture Model, *Afrika Statistika*, **11**, 1041-1059.

MacQueen, J. (1967), Some Methods for Classification and Analysis of Multivariate Observations, In*Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

Mangasarian, O. L., Street, W. N. and Wolberg, W. H. (1995), Breast Cancer Diagnosis and Prognosis via Linear Programming, *Operations Research*, **43**, 570-577.

McLachlan, G., and Peel, D. (2004), *Finite Mixture Models*, Wiley Series in Probabilities and Statistics, New York, NY: Wiley-Interscience.

Meinshausen, N. and Bühlmann, P. (2006), High-dimensional Graphs and Variable Selection with the Lasso, *The Annals of Statistics*, **34**, 1436-1462.

Mukherjee, S. and Hill, S. M. (2011), Network Clustering: Probing Biological Heterogeneity by Sparse Graphical Models, *Bioinformatics*, **27**, 994-1000.

Rand, W. M. (1971), Objective Criteria for the Evaluation of Clustering Methods, *Journal of the American Statistical Association*, **66**, 846-850.

Städler, N., Bühlmann, P. and Van de Geer, S. (2010), ℓ_1-Penalization for Mixture Regression Models, *TEST*, **19**, 209-256.

Yuan, M. and Lin, Y. (2007), Model Selection and Estimation in the Gaussian Graphical Model, *Biometrika*, **94**, 19-35.

Zhou, H., Pan, W. and Shen, X. (2009), Penalized Model-based Clustering with Unconstrained Covariance Matrices, *Electronic Journal of Statistics*, **3**, 1473-1496.