مجله علوم آماری، پاییز و زمستان ۱۳۹۶
جلد ۱۱، شماره ۲، ص ۲۶۳ – ۲۸۴
DOI: 10.18869/acadpub.jss.11.2.263

خوشه‌بندی پروفایل‌های طولی با مدل‌های اثرات آمیخته ناپارامتری و نیمه‌پارامتری

میثم تسلی‌زاده خمس[۱]، زهرا رضایی قهرودی[۲]

[۱] گروه آمار، دانشگاه تربیت مدرس

[۲] پژوهشکده آمار

تاریخ دریافت: ۱۳۹۵/۱/۲۲ تاریخ آخرین بازنگری: ۱۳۹۶/۹/۲۵

چکیده: روش‌های متعددی برای خوشه‌بندی داده‌های بیان ژن دوره‌ای زمانی وجود دارد ولی محدودیت‌هایی برای این روش‌ها وجود دارد که از جمله آن‌ها می‌توان به عدم در نظر گرفتن همبستگی در طول زمان و زمان‌بر بودن محاسبات اشاره داشت. در این مقاله با معرفی مدل‌های اثرات آمیخته‌ی ناپارامتری و نیمه‌پارامتری، این همبستگی در طول زمان در نظر گرفته شده و با استفاده از اسپلاین تاوانیده، حجم محاسبات به طور چشم‌گیری کاهش یافته است. در پایان با استفاده از مطالعه شبیه‌سازی عملکرد روش پیشنهادی با روش‌های قبلی مقایسه و با استفاده از ملاک BIC، مدل مناسب‌تر انتخاب و تحلیل می‌شود. همچنین روش پیشنهادی در یک مثال کاربردی داده‌های بیان ژن دوره‌ای زمانی ارائه شده است.

واژه‌های کلیدی: پروفایل‌های طولی، هموارسازی اسپلاین، اسپلاین تاوانیده، مدل اثرات آمیخته خطی، خوشه‌بندی مدل پایه، بیان ژن.

۱ مقدمه

به دلیل توسعه جمع‌آوری و ثبت داده‌هایی با وابستگی طولی و کاربرد مطالعات طولی در حوزه‌های مختلف، روش‌های مختلفی برای تحلیل این نوع داده‌ها توسعه و گسترش پیدا کرده است. یکی از حوزه‌های مطالعاتی داده‌های طولی در پزشکی است. کاربردهای مطالعات طولی در مباحث ژنتیک و بیوانفورماتیک نیز وجود دارد و داده‌های بیان ژن دوره‌ای زمانی، از شاخص‌ترین داده‌ها در این زمینه است که دارای ویژگی طولی

آدرس الکترونیکی نویسنده مسئول مقاله: زهرا رضایی قهرودی، z_rezaei@srtc.ac.ir

کد موضوع‌بندی ریاضی (۲۰۱۰): 62G08، 62H30

است. از آنجا که خوشه‌بندی داده‌های بیان ژن بافت‌های سلولی به خوشه‌های همگن به منظور بررسی دامنه وسیعی از فرایندهای زیست‌شناختی، ضروری است، علاقه زیادی به خوشه‌بندی داده‌های بیان ژن برای به دست آوردن خوشه‌هایی با بیان ژن مشابه وجود دارد. آیزن و همکاران (۱۹۹۸) برای اولین بار با استفاده از روش سلسله مراتبی به خوشه‌بندی داده‌های بیان ژن پرداختند. یکی از موفق‌ترین نمونه‌های کاربرد خوشه‌بندی در داده‌های بیان ژن، خوشه‌بندی داده‌های بیان ژن بیماران سرطانی توسط علیزاده و همکاران (۲۰۰۰) است. با وجود انجام فعالیت‌های زیادی در این حوزه، هنوز خوشه‌بندی این نوع داده‌ها برای تعیین خوشه‌های درست و دقیق چالش برانگیز است.

از طرفی داده‌های بیان ژن دوره‌ای زمانی یا داده‌های با ابعاد بالا می‌توانند در قالب‌های مختلفی گنجانده شوند که داده‌های طولی، نمونه‌ای از این داده‌ها هستند. لذا خوشه‌بندی این نوع داده‌ها یعنی کاهش بعد داده‌ها و پیدا کردن گروه‌های مشابه از مشاهدات در طول زمان از اهمیت به‌سزایی برخوردار است. استفاده از روش‌های مناسب خوشه‌بندی، برای خوشه‌بندی داده‌های بیان ژن دوره‌ای زمانی برای تعیین گروه‌هایی از ژن‌ها که در طول زمان رفتار مشابهی دارند، از موضوعات مهم به شمار می‌رود. روش‌هایی که پیش از این برای خوشه‌بندی داده‌های بیان ژن دوره‌ای زمانی مورد استفاده قرار گرفته‌اند، با محدودیت‌هایی مانند عدم درنظر گرفتن ساختار همبستگی در طول زمان، حجم بالای محاسبات و زمان بر بودن محاسبات روبرو هستند. از طرفی به دلیل اینکه ماهیت داده‌های بیان ژن دوره‌ای زمانی در طول زمان یک فرایند پیوسته است، برای نمایش این داده‌ها در طول زمان بایستی از یک تابع پیوسته استفاده کرد که استفاده از روش‌های هموارسازی می‌تواند مفید باشد. همچنین با توجه به طبیعت این داده‌ها که یک نوع وابستگی در طول زمان بین آنها وجود دارد و نیز با خطای تصادفی و غیرقابل اندازه‌گیری در طول زمان مواجه‌ایم، بنابراین برای تحلیل این نوع داده‌ها از مدل اثرات آمیخته و اسپلاین تاوانیده استفاده می‌شود.

نخستین بار، بارجوزف و همکاران (۲۰۰۳) با استفاده از مدل اسپلاین مکعبی به خوشه‌بندی داده‌های بیان ژن دوره‌ای زمانی پرداختند. چن (۲۰۰۹) با استفاده از روش خوشه‌بندی خودسازمانده[1] (SOM) روش جدیدی را که مبتنی بر نقشه‌های خودسازمانده است، برای خوشه‌بندی داده‌های طولی معرفی کرد. همچنین جنولینی و فالیسارد (۲۰۱۰) با معرفی روش $K-$میانگین طولی[2] (KML) به خوشه‌بندی داده‌های طولی با روش $K-$میانگین پرداختند.

از طرفی مدل‌های اثرات آمیخته که ابزار قوی برای تحلیل مطالعات طولی است را می‌توان به صورت پارامتری (خطی و غیرخطی)، ناپارامتری و نیمه‌پارامتری بیان کرد. هارویل (۱۹۷۷، ۱۹۷۶) و لیرد و

[1] Self-Organizing Map
[2] K-mean Longitudinal

ویر (۱۹۸۲) برای اولین بار مدل اثرات آمیخته خطی پارامتری را معرفی کردند. مدل‌های اثرات آمیخته غیرخطی که رابطه بین متغیر پاسخ و کمکی را از طریق مدل غیرخطی بیان می‌کند، اولین بار توسط داویدیان و گیلتینان (۱۹۹۵) معرفی شد. در مدل اثرات آمیخته پارامتری فرض بر این است که شکل مدل برای داده‌ها به صورت خطی یا غیرخطی مشخص است که گاهی اوقات گذاشتن این شرط بر داده‌های طولی نامعتبر است. از این رو مدل‌های دیگری برای مدل‌بندی داده‌های طولی تحت عنوان مدل اثرات آمیخته ناپارامتری (ژانگ و وو، ۲۰۰۶) معرفی شده است. مدل‌بندی اثرات ثابت و تصادفی ناپارامتری در مدل ناپارامتری به روش‌های مختلفی از جمله روش اسپلاین رگرسیونی، اسپلاین هموارساز (واهبا، ۱۹۹۰) و اسپلاین جریمه‌ای (روپرت و همکاران، ۲۰۰۳) صورت گرفته است. در مدل‌های نیمه‌پارامتری نیز از مؤلفه‌های پارامتری برای عامل‌های مهم مدل و از مؤلفه‌های نیمه‌پارامتری برای عامل‌های با اهمیت کمتر استفاده می‌شود (فایفر، ۲۰۰۴). فایفر (۲۰۰۴) به خوشه‌بندی پروفایل‌های طولی با استفاده از مدل اثرات آمیخته نیمه‌پارامتری و $B-$اسپلاین پرداخته است. حجم بالای محاسبات و مدت زمان طولانی محاسبات از نقاط ضعف این روش برای داده‌های بیان ژن دوره‌ای زمانی با حجم بالا است. در فایفر (۲۰۰۴) پارامترهای مدل با استفاده از الگوریتم EM برآورد شده است که وقتی با حجم بالای داده‌های بیان ژن دوره‌ای زمانی روبرو هستیم، به همگرایی دست نمی‌یابد. همچنین در این مقاله روش تعیین تعداد خوشه براساس مقادیر لگاریتم تابع درستنمایی بوده است که به برخی مشکلات از جمله عدم پایداری تابع درستنمایی در روش نیمه‌پارامتری و مواجهه با مشکلاتی در تعیین تعداد خوشه مناسب براساس معیارهای BIC و AIC اشاره شده است. کوفی و همکاران (۲۰۱۴) روش جدیدی را معرفی کردند که در آن با استفاده از مدل اثرات آمیخته خطی و هموارسازی اسپلاین تاوانیده که به مدل‌های اثرات آمیخته ناپارامتری معروف‌اند، به خوشه‌بندی داده‌های طولی پرداختند. در این روش ابتدا هرگونه خطای قابل اندازه‌گیری در مشاهدات بوسیله هموارسازی از بین می‌رود و تأثیرات غیرقابل اندازه‌گیری و وابستگی در طول زمان نیز از طریق اثرات تصادفی در مدل در نظر گرفته می‌شود، سپس داده‌های بیان ژن با استفاده از روش مبتنی بر مدل، خوشه‌بندی می‌شود. استواری و سرعت محاسبات بالاتر برای داده‌های با بعد بالا به دلیل استفاده از الگوریتم RCEM به جای لگوریتم EM و تعیین تعداد دقیق خوشه براساس معیار BIC از جمله مزیت‌های روش پیشنهادی کوفی و همکاران (۲۰۱۴) در مقایسه با فایفر (۲۰۰۴) است.

با توجه به اینکه در بسیاری مطالعات متغیرهای کمکی بر پاسخ‌های مورد بررسی تاثیر زیادی دارند، هدف این مقاله خوشه‌بندی داده‌های بیان ژن دوره‌ای زمانی با استفاده از مدل‌های اثرات آمیخته نیمه‌پارامتری و مقایسه آن با مدل‌های مختلف ناپارامتری است. بعلاوه با استفاده از معیار BIC و مدت زمان تحلیل

داده‌ها، خوشه‌بندی داده‌های طولی با استفاده از مدل نیمه‌پارامتری و مدل‌های مختلف ناپارامتری مقایسه شده است که نتایج آن نشان‌دهنده تشخیص درست مدل است. از این رو در بخش ۲ روش‌های هموارسازی معرفی می‌شوند. سپس در بخش ۳ مدل‌های رگرسیون آمیخته پارامتری، ناپارامتری و نیمه‌پارامتری معرفی می‌شوند. در بخش ۴ نیز خوشه‌بندی با استفاده از مدل اثرات آمیخته ناپارامتری و نیمه‌پارامتری داده‌های طولی معرفی می‌شود. در بخش ۵ نتایج شبیه‌سازی و مقایسه عملکرد روش‌های ناپارامتری و نیمه‌پارامتری برای داده‌های بیان ژن دوره‌ای زمانی ارائه شده است. همچنین در بخش ۶، روش پیشنهادی برای تحلیل داده‌های بیان ژن دوره‌ای زمانی بکار گرفته شده است و در آخر به بحث و نتیجه‌گیری پرداخته شده است.

۲ هموارسازی

داده‌های بیان ژن دوره‌ای زمانی همانند مطالعات طولی، در یک دوره‌ی زمانی اندازه‌گیری می‌شوند. اگر Y_{ij} نشان دهنده مقدار بیان ژن مربوط به ژنiاُم در زمان jاُم باشد $(i = ۱, \ldots, N,\ j = ۱, \ldots, n_i)$ که در آن N تعداد کل ژن‌ها و n_i تعداد تکرارها برای ژن iاُم است می‌توان رفتار یک ژن را در طول زمان با استفاده از یک منحنی هموار نشان داد به‌طوریکه داده‌ها شکل منحنی پیوسته را نشان می‌دهد. بنابراین می‌توان داده‌ها را با استفاده از توزیع نرمال با میانگین $\mu(t_i)$ مدل‌بندی کرد. مدل رگرسیون ناپارامتری برای داده‌ها به‌صورت

$$Y_{ij} = h(t_{ij}) + \varepsilon_{ij}$$

است که در آن $h(t)$ تابعی هموار در زمان t است که لازم است براورد شود. روش‌های زیادی برای برآورد تابع $h(t)$ وجود دارد. در این روش‌ها با استفاده از ترکیب خطی از K تابع پایه‌ای $\{\phi_۱(t), \ldots, \phi_K(t)\}$، می‌توان تابع $h(t)$ را به صورت

$$h(t) = \sum_{j=۱}^{K} \beta_j \phi_j(t) \qquad (۱)$$

نوشت که در آن برای $j = ۱, \ldots, K$، $\phi_j(t)$، توابع پایه‌ای هستند که بایستی برآورد شوند.

۱.۲ اسپلاین رگرسیونی

اسپلاین رگرسیونی یکی از روش‌های هموارسازی است که در آن از توابع پایه‌ای توانی بریده شده از درجه‌ی k به‌صورت

$$۱, t, \ldots, t^k, (t-\xi_۱)_+^k, \ldots, (t-\xi_K)_+^k$$

استفاده می‌شود که در آن $\xi_۱, \ldots, \xi_K$ گره نامیده می‌شود و $(t-\xi_j)_+^k = \max(۰, (t-\xi_j)_+^k)$. با استفاده از این توابع پایه‌ای، رابطه (۱) به‌صورت

$$h(t) = \sum_{s=۰}^{k} \beta_s t^s + \sum_{j=۱}^{K} \beta_{k+j}(t-\xi_j)_+^k \qquad (۲)$$

بیان می‌شود که در آن $\beta = (\beta_۰, \ldots, \beta_{k+K})$ بردار ضرایب رگرسیونی هستند و می‌توان آن را به صورت

$$h(t) = \Phi_p(t)'\beta \qquad (۳)$$

بازنویسی کرد که در آن $\Phi_p(t) = (۱, t, \ldots, t^k, (t-\xi_۱)_+^k, \ldots, (t-\xi_K)_+^k)$ و $p = k+K+۱$.
با استفاده از (۲) مدل اسپلاین رگرسیونی را می‌توان به صورت ماتریسی $Y = T\beta + \varepsilon$ ارائه کرد که در آن $T = (\Phi_p(t_۱), \ldots, \Phi_p(t_N))'$ ماتریس طرح $N \times (k+K+۱)$ بعدی است. ضرایب در (۳) بر اساس ملاک کمترین توان‌های دوم عادی

$$RSS = \sum_{i=۱}^{N}(y_i - h(t_i))^۲ = ||Y - T\beta||^۲ \qquad (۴)$$

به صورت $\hat{\beta} = (T'T)^{-۱}T'Y$ برآورد می‌شود و $\hat{Y} = T\hat{\beta} = T(T'T)^{-۱}T'Y$.

۲.۲ اسپلاین رویه نازک

وقتی k فرد باشد، بر اساس توابع پایه‌ای شعاعی[۳] $۱, t, \ldots, t^k, |t-\xi_۱|_+^k, \ldots, |t-\xi_K|_+^k$ اسپلاین رویه نازک به صورت $h(t) = \beta_۰ + \beta_۱ t + \sum_{j=۱}^{K} \beta_{۱+j}(t-\xi_j)_+^۳$ تعریف می‌شود (شن، ۲۰۱۱).

[3] Radial basis function

۳.۲ اسپلاین تاوانیده

از آنجا که تعداد گره‌ها نامعلوم است، معمولاً تعداد گره‌ها را با تعداد مشاهدات برابر می‌گیرند که در این صورت با بیش برازندگی مواجه خواهیم شد. برای کنترل و جلوگیری از بیش برازندگی، با اضافه کردن یک عبارت جریمه به صورت $\int (D^۲h(t))^۲dt$ به برآورد پارامترهای مدل، از ملاک مجموع توان‌های دوم مانده‌های جریمه‌ای[۴] (PRSS) به صورت

$$PRSS = ||Y - T\beta||^۲ + \lambda \int (D^۲h(t))^۲dt \qquad (۵)$$

استفاده می‌شود که در آن D ماتریس $(k + K + ۱) \times (k + K + ۱)$بعدی به صورت

$$D = \begin{bmatrix} O_{(k+۱)\times(k+۱)} & O_{(k+۱)\times K} \\ O_{K\times(k+۱)} & I_{K\times K} \end{bmatrix}$$

و $\lambda > ۰$ پارامتر همواری است که با روش اعتبارسنجی متقابل بدست می‌آید. همچنین $D^۲h(t)$ نشان‌دهنده مشتق دوم $h(t)$ است و میزان انحنای منحنی $h(t)$ را نشان می‌دهد.

یکی از بزرگترین معایب اسپلاین هموارساز در وابستگی به تعداد گره‌ها است به طوری که با افزودن گره‌های جدید با مشکلاتی از قبیل بیش‌برازندگی و محاسبه انتگرال‌های پیچیده مواجه می‌شویم. از این‌رو از روشی تحت عنوان هموارسازی اسپلاین تاوانیده یا $P-$اسپلاین استفاده می‌شود و بیش‌برازندگی با استفاده از محدودیت بر روی ضرایب جریمه‌ای به صورت $\sum_{j=۱}^{K} \beta_{k+j}^۲ < C$ گزارش می‌شود، که در آن C یک ثابت عددی است. روپرت (۲۰۰۲) روشی ارائه کرد، که در آن تعداد گره‌ها برابر با $K = \max(۵, \min(\frac{N}{۴}, ۳۵))$ و موقعیت گره‌ها $\xi_j = \frac{j+۱}{K+۲}$, $j = ۱, \ldots, K$ اختیار می‌شوند. برای برازش اسپلاین تاوانیده از ملاک

$$PRSS(\beta, \lambda) = ||Y - T\beta||^۲ + \lambda \beta^T D \beta \qquad (۶)$$

استفاده می‌شود. لازم به ذکر است که $(\beta_{k+۱}, \ldots, \beta_{k+K})$ ضرایب تاوانیده‌اند به‌طوری که

[4] Penalized Residual Sum of Squares

$\beta' D\beta = \sum_{j=1}^{K} \beta_{k+j}$. از اینرو عبارت جریمه هموارسازی در اسپلاین تاوانیده به جای انتگرال، یک مجموع گسسته خواهد شد (گیرین و سیلورمن ۱۹۹۴، روپرت و همکاران، ۲۰۰۳).

۳ مدل‌بندی داده‌های بیان ژن با مدل اثرات آمیخته

از آنجا که مدل‌های اثرات آمیخته خطی یک ابزار قوی برای تحلیل مطالعات طولی است، در این بخش به مطالعه این مدل‌ها در مطالعات طولی پرداخته می‌شود و ساختار همبستگی بین پاسخ‌های آزمودنی‌ها در طول زمان و یا اثرات تثبیت نشده هر آزمودنی به صورت اثر تصادفی در مدل وارد می‌شود.

۱.۳ مدل اثرات آمیخته ناپارامتری

مجموعه داده‌های طولی (t_{ij}, y_{ij}) برای $i = 1, \ldots, N$، $j = 1, \ldots, n_i$ را در نظر بگیرید که در آن y_{ij} پاسخ مشاهده شده در زمان t_{ij} و N تعداد آزمودنی‌ها باشد. ژانگ و وو (۲۰۰۶) مدل اثرات آمیخته ناپارامتری را به صورت

$$y_{ij}(t) = S(t_{ij}) + v_i(t_{ij}) + \varepsilon_{ij}, \qquad i = 1, \ldots, N, \;\; j = 1, \ldots, n_i \qquad (۷)$$

معرفی کردند که در آن $S(t_{ij})$ تابع اثرات ثابت ناپارامتری (مدل تابع میانگین جامعه) و $v_i(t_{ij})$ تابع اثرات تصادفی ناپارامتری (مدل انحراف آزمودنی iاُم از تابع میانگین جامعه) برای فرد iاُم است. ε_{ij} خطاهای قابل اندازه‌گیری است که توسط $v_i(t_{ij})$ قابل بیان نیست. برای مدل‌بندی مولفه‌های $S(t_{ij})$ و $v_i(t_{ij})$ از روش‌های ناپارامتری از جمله روش‌های اسپلاین رگرسیونی، اسپلاین رویه نازک، اسپلاین هموارساز (واهبا، ۱۹۹۰) و اسپلاین تاوانیده (آیلرز و مارکس، ۱۹۹۶؛ روپرت و همکاران، ۲۰۰۳) استفاده نمود.

۲.۳ مدل اثرات آمیخته نیمه‌پارامتری

در مدل‌های اثرات آمیخته نیمه‌پارامتری از مولفه‌های پارامتری اغلب برای فاکتورهای مهم مدل استفاده می‌شود و از مولفه‌های ناپارامتری برای فاکتورهای با اهمیت کمتر استفاده می‌شود. از طرفی به منظور در نظر گرفتن همبستگی بین آزمودنی‌ها، اثرات تصادفی نیز در مدل اضافه می‌شود که می‌تواند بدون وابستگی

به زمان t_{ij} و یا وابستگی به متغیرهای کمکی به‌صورت

$$Y_{ij}(t) = x_{ij}^T\alpha + S(t_{ij}) + b_i + \varepsilon_{ij}, \qquad i = ۱, \ldots, N, \;\; j = ۱, \ldots, n_i \qquad (۸)$$

در مدل ظاهر شود، به گونه‌ای که برای متغیر پاسخ Y و متغیرهای کمکی x و t داریم:

$$E[Y|x, t] = x^T\alpha + S(t_{ij})$$

حالت تعمیم‌یافته این مدل در نظر گرفتن وابستگی مولفه‌ی اثر تصادفی به زمان t_{ij} از طریق یک فرایند هموارسازی $v_i(t_{ij})$ و وابستگی خطی به متغیر کمکی دیگر از طریق بردار $z_{ij} = [z_{۱ij}, \ldots, z_{qij}]^T$ بیان می‌شود که در این صورت مدل اثرات آمیخته نیمه‌پارامتری به صورت

$$Y_{ij}(t) = x_{ij}^T\alpha + S(t_{ij}) + z_{ij}^T u_i + v_i(t_{ij}) + \varepsilon_{ij}, \;\; i = ۱, \ldots, N, \;\; j = ۱, \ldots, n_i \;\; (۹)$$

تعریف می‌شود که در آن Y_{ij} متغیر پاسخ آزمودنی iاُم در زمان t_{ij} و $x_{ij} = [x_{۱ij}, \ldots, x_{pij}]^T$ متغیرهای کمکی مشاهده شده در زمان t_{ij} است که تعداد آن‌ها p تا است. $\alpha = [\alpha_۱, \ldots, \alpha_p]^T$ بردار ضرایب متغیرهای تبیینی، $S(\cdot)$ تابعی هموار از زمان، $v_i(\cdot)$ فرایند هموارسازی که وابستگی مولفه‌های اثرهای تصادفی به زمان t_{ij} را در نظر می‌گیرد و $z_{ij} = [z_{۱ij}, \ldots, z_{qij}]^T$ بردار متغیرهای کمکی است که وابستگی خطی اثرهای تصادفی با متغیرهای کمکی دیگر را بیان می‌کند. همچنین $u_i = [u_{i۱}, \ldots, u_{iq}]^T$ ضرایب متغیرهای تبیینی بردار z_{ij} است و برای $i = ۱, \ldots, N$، $v_i(t) \sim GP(۰, \gamma)$، $u_i \sim N(۰, G_u)$ و $\varepsilon_i = (\varepsilon_{i۱}, \ldots, \varepsilon_{in_i})^T \sim N(۰, R_i)$ است که ε_{ij} میزان خطای اندازه‌گیری برای مشاهده iاُم در زمان jاُم است و $GP(۰, \gamma)$ یک فرایند گاوسی با میانگین صفر و تابع کواریانس $\gamma(s, t)$ است. در این مدل، $x_{ij}^T\alpha$ اثرات ثابت پارامتری، $S(t_{ij})$ اثرات ثابت ناپارامتری، $z_{ij}^T u_i$ اثرات تصادفی پارامتری و $v_i(t_{ij})$ اثرات تصادفی ناپارامتری هستند.

۴ خوشه‌بندی پروفایل‌های طولی با مدل اثرات آمیخته نیمه‌پارامتری

مدل‌های آمیخته متناهی ابزار مناسبی برای تحلیل خوشه‌ای فراهم می‌کند به گونه‌ای که در روش خوشه‌بندی مبتنی بر مدل آمیخته، فرض بر این است که داده‌ها از یک جامعه آمیخته با تعداد مشخص خوشه و

نسبت‌های آمیخته متفاوت آمده‌اند. از طرفی با استفاده از روش‌های هموارسازی می‌توان تابع $S(t)$ را برای یک ژن در طول زمان مدل‌سازی کرد و از این روش برای مدل‌بندی خوشه‌های بیان ژن بر اساس تابع میانگین هر خوشه استفاده کرد. فرض کنید آزمودنی iاُم در خوشه gاُم باشد. در این‌صورت برای مدل‌بندی مشاهده‌ی i در خوشه‌ی gاُم، مدل اثرات آمیخته نیمه‌پارامتری به‌صورت

$$y_{ij}(t) = x_{ij}^T\alpha + S_g(t_{ij}) + b_i + \varepsilon_{ij}, \qquad i = ۱, \ldots, N_g, \;\; j = ۱, \ldots, n_i \qquad (۱۰)$$

است، که در آن $S_g(t_{ij})$ تابع میانگین برای خوشه‌ی gاُم، $x_{ij} = [x_{۱ij}, \ldots, x_{pij}]^T$ متغیرهای کمکی مشاهده شده در زمان t_{ij} است. $\alpha = [\alpha_۱, \ldots, \alpha_p]^T$ بردار ضرایب متغیرهای تبیینی، $\varepsilon_{ij} \sim N(۰, \sigma^۲_{\varepsilon g})$، N_g تعداد مشاهدات در خوشه‌ی gاُم و n_i تعداد تکرارها برای آزمودنی iاُم است. اگر تابع میانگین $S_g(t_{ij})$ را با استفاده از تابع اسپلاین تاوانیده مدل‌بندی کنیم، مدل به صورت

$$Y_g = X_g^T\alpha_g + \underbrace{T_{g,s}\beta_{g,s} + Z_{g,s}u_{g,s}}_{s_g(t)} + Z_{g,b}b_g + \varepsilon_g \qquad (۱۱)$$

است، که در آن $Y_g = (y_۱^T, \ldots, y_{N_g}^T)$، $y_i = (y_{i۱}, \ldots, y_{in_i})^T$، $X_g = (x_۱^T, \ldots, x_{N_g}^T)$، $x_i = (x_{i۱}, \ldots, x_{in_i})^T$، $T_{g,s} = (T_{۱,s}, \ldots, T_{N_g,s})^T$، $Z_{g,s} = (Z_{۱,s}, \ldots, Z_{N_g,s})^T$،

$$T_{i,s} = \begin{bmatrix} ۱ & t_{i۱} & \ldots & t_{i۱}^k \\ ۱ & t_{i۲} & \ldots & t_{i۲}^k \\ \vdots & \vdots & \ddots & \vdots \\ ۱ & t_{in_i} & \ldots & t_{in_i}^k \end{bmatrix}, \qquad Z_{i,s} = \begin{bmatrix} (t_{i۱} - \xi_۱)_+^k & \ldots & (t_{i۱} - \xi_K)_+^k \\ (t_{i۲} - \xi_۱)_+^k & \ldots & (t_{i۲} - \xi_K)_+^k \\ \vdots & \ddots & \vdots \\ (t_{in_i} - \xi_۱)_+^k & \ldots & (t_{in_i} - \xi_K)_+^k \end{bmatrix}$$

و $Z_{g,b} = diag(Z_{۱,b}, \ldots, Z_{N_g,b})$ و $Z_{۱,b} = (۱, \ldots, ۱)^T$ است. مدل نیمه‌پارامتری (۱۱) برای ژن iاُم در خوشه gاُم به صورت

$$Y_i = X_i\alpha_g + \underbrace{T_{i,s}\beta_{g,s} + Z_{i,s}u_{g,s}}_{s_g(t_i)} + Z_{i,b}b_i + \varepsilon_i$$

نوشته می‌شود که در آن $\alpha_g = [\alpha_{g0}, \ldots, \alpha_{gp}]$ است. برای خوشه‌بندی با روش مدل پایه فرض بر این است که توزیع ژن iاُم در خوشه gاُم به شرط اثر تصادفی که برای هموارسازی است $(u_{g,s})$ نرمال $y_i|u_{g,s} \sim N(X_i\alpha_g + T_{i,s}\beta_{g,s} + Z_{i,s}u_{g,s}, V_{ig})$ است، که در آن

$$V_{ig} = \sigma^2_{bg} Z_{i,b} Z^T_{i,b} + \sigma^2_{\varepsilon g} I_{n_i \times n_i}$$

متناظر با بلوک $n_i \times n_i$ و ماتریس کواریانس $V_g = \sigma^2_{bg} Z_{g,b} Z^T_{g,b} + \sigma^2_{\varepsilon g} I$ مربوط به خوشه gاُم است. این عبارت نشان می‌دهد که $\sigma^2_{\varepsilon g}$ و σ^2_{bg} برای همه ژن‌ها در خوشه gاُم یکسان هستند. ماتریس V_{ig} به صورت

$$V_{ig} = \sigma^2_{bg} \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix} + \sigma^2_{\varepsilon g} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \ldots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2_{bg} + \sigma^2_{\varepsilon g} & \sigma^2_{bg} & \ldots & \sigma^2_{bg} \\ \sigma^2_{bg} & \sigma^2_{bg} + \sigma^2_{\varepsilon g} & \ldots & \sigma^2_{bg} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{bg} & \sigma^2_{bg} & \ldots & \sigma^2_{bg} + \sigma^2_{\varepsilon g} \end{bmatrix}$$

است. برای برازش هموارساز اسپلاین تاوانیده، ملاک (۶) برای خوشه gاُم به صورت

$$PLS(\beta, \lambda) = ||Y_g - T_{g,s}\beta_{g,s} - Z_{g,s}u_{g,s}||^2 + \lambda_g ||u_{g,s}||^2 \qquad (۱۲)$$

است. با کمینه کردن ملاک (۱۲) با استفاده از معیار بهترین پیشگوی خطی نااریب[5] (BLUP) برآورد اسپلاین تاوانیده به صورت

$$\hat{Y}_g = C_g (C_g^T C_g + \lambda_g D)^{-1} C_g^T Y_g$$

[5] Best Linear Unbiased Prediction

خواهد بود که در آن $C_g = [T_{g,s}, Z_{g,s}]$ و $\lambda_g = \frac{\sigma^2_{\varepsilon g}}{\sigma^2_{ug}}$ (روپرت و همکاران ۲۰۰۳). در عمل، اعضای خوشه نامعلوم هستند و فرض بر این است که y_i از توزیع آمیخته

$$f_Y(y_i;\Theta) = \sum_{g=1}^{G} \pi_g f_g(y_i;\theta_g)$$

بدست می‌آید که در آن $f_g(\cdot)$ تابع چگالی مربوط به خوشه gاُم با پارامتر $\theta_g = (\alpha_g, \beta_{g,s}, \sigma^2_{ug}, \sigma^2_{bg}, \sigma^2_{\varepsilon g})$، $\Theta = (\theta_1^T, \ldots, \theta_G^T, \pi_1, \ldots, \pi_{G-1}, \pi_G)^T$ و $\pi_1, \ldots, \pi_G$ نسبت‌های آمیخته هستند که احتمال تعلق مشاهدات به خوشه‌ها را نشان می‌دهند به طوری که $\sum_{g=1}^{G} \pi_g = 1$. بنابراین برای برآورد منحنی میانگین همواری در هر خوشه، به شرط اثر تصادفی $(u_{g,s})$ تابع چگالی آزمودنی iاُم به صورت

$$\begin{aligned}\prod_{i=1}^{N_T} f_{Y_i}(y_i|u_{g,s};\Theta) &= \prod_{i=1}^{N_T}\sum_{g=1}^{G} \pi_g N(X_i\alpha_g + S_g(t_i), V_{ig}) \\ &= \prod_{i=1}^{N_T}\sum_{g=1}^{G} \pi_g N(X_i\alpha_g + T_{i,s}\beta_{g,s} + Z_{i,s}u_{g,s}, V_{ig}) \quad (۱۳)\end{aligned}$$

است، که در آن $N_T = \sum_{g=1}^{G} N_g$ تعداد کل آزمودنی‌ها است. تابع لگاریتم درستنمایی برای مدل آمیخته عبارتست از:

$$\log L = \sum_{i=1}^{N_T} \log \sum_{g=1}^{G} \pi_g N(X_i\alpha_g + T_{i,s}\beta_{g,s} + Z_{i,s}u_{g,s}, V_{ig}) \qquad (۱۴)$$

که برای ماکسیمم‌سازی آن با استفاده از الگوریتم EM نیاز به معرفی متغیر نشانگر به صورت

$$z_{ig} = \begin{cases} 1 & \text{اگر ژن } i\text{ام به خوشه } g\text{ام متعلق باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

است. در این صورت تابع لگاریتم درستنمایی (۱۴) عبارت خواهد شد از

$$\begin{aligned}\log L &= \sum_{i=1}^{N_T}\{\sum_{g=1}^{G} z_{ig}[\log \pi_g + \log N(X_i\alpha_g + T_{i,s}\beta_{g,s} + Z_{i,s}u_{g,s}, V_{ig})]\} \\ &= \sum_{i=1}^{N_T}\{\sum_{g=1}^{G} z_{ig}[\log \pi_g + \log N(\mu_g(t_i), V_{ig})]\}. \qquad (۱۵)\end{aligned}$$

۱.۴ الگوریتم EM

الگوریتم EM بین دو مرحله امید ریاضی (مرحله E) و مرحله ماکسیمم‌سازی (مرحله M) تکرار می‌شود. در تکرار $(r+1)$اُم، مرحله امید ریاضی الگوریتم EM به محاسبه احتمال پسین

$$\begin{aligned}P(z_{ig}) &= 1|y_i, S_g^{(r)}(t), \hat{V}_g^{(r)}, \hat{\alpha}_g^{(r)}) = \hat{\tau}_{ig}^{(r+1)} \\ &= \frac{\hat{\pi}_g^{(r)} N(X_i\hat{\alpha}_g^{(r)} + \hat{S}_g^{(r)}(t_i), \hat{V}_{ig}^{(r)})}{\sum_{h=1}^{G} \hat{\pi}_h^{(r)} N(X_i\hat{\alpha}_h^{(r)} + \hat{S}_h^{(r)}(t_i), \hat{V}_{ih}^{(r)})} \qquad (۱۶)\end{aligned}$$

یعنی میزان تعلق آزمودنی iاُم به خوشه‌ی gاُم، می‌پردازد که در آن $\hat{S}_g^{(r)}(t_i)$ برآورد منحنی میانگین در خوشه‌ی gاُم در تکرار r است که برای n_i نقطه زمانی برای آزمودنی iاُم بدست آمده است و به صورت

$$\hat{S}_g(t_i) = \begin{bmatrix} \hat{\alpha}_{g0} + \cdots + \hat{\alpha}_{gp}x_{i1p} + \hat{\beta}_{g0} + \cdots + \hat{\beta}_{gk}t_{i1}^k + \hat{\beta}_{g,k+1}(t_{i1} - \xi_1)_+^k + \cdots + \hat{\beta}_{g,k+K}(t_{i1} - \xi_K)_+^k \\ \hat{\alpha}_{g0} + \cdots + \hat{\alpha}_{gp}x_{i2p} + \hat{\beta}_{g0} + \cdots + \hat{\beta}_{gk}t_{i2}^k + \hat{\beta}_{g,k+1}(t_{i2} - \xi_1)_+^k + \cdots + \hat{\beta}_{g,k+K}(t_{i2} - \xi_K)_+^k \\ \vdots \\ \hat{\alpha}_{g0} + \cdots + \hat{\alpha}_{gp}x_{in_ip} + \hat{\beta}_{g0} + \cdots + \hat{\beta}_{gk}t_{in_i}^k + \hat{\beta}_{g,k+1}(t_{in_i} - \xi_1)_+^k + \cdots + \hat{\beta}_{g,k+K}(t_{in_i} - \xi_K)_+^k \end{bmatrix}$$

است و $x_i\hat{\alpha}_g^{(r)}$ برآورد اثرات ثابت پارامتری در خوشه gاُم در تکرار rاُم و $\hat{V}_{ig}^{(r)}$ ماتریس $n_i \times n_i$تایی برآورد ماتریس واریانس کواریانس $\hat{V}_g^{(r)}$ است. پس از بدست آوردن برآوردگرهای $\hat{\tau}_{ig}$ در تکرار $(r+1)$اُم، با استفاده از (۱۶)، نسبت‌های آمیختن $\hat{\pi}_g^{(r+1)} = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{g=1}^{G}\hat{\tau}_{ig}^{(r+1)}$ برای هر خوشه برآورد می‌شود که در آن $\sum_{i=1}^{N} n_i$ اندازه تمام مشاهدات است. در مرحله ماکسیمم‌سازی، برآوردهای $\hat{S}_g^{(r+1)}(t)$ و $\hat{V}_g^{(r+1)}$ و $\hat{\alpha}_g^{(r+1)}$ با برازش مدل (۱۱) و استفاده از $\hat{\tau}_{ig}^{(r+1)} = (\hat{\tau}_{1g}, \ldots, \hat{\tau}_{Ng})'$ به

عنوان وزن، به هنگام می‌شوند. این رویه بین مراحل E و M تا جایی که به همگرایی برسد، ادامه می‌یابد.
مرحله امیدگیری: مرحله امیدگیری در الگوریتم EM، در ارتباط با محاسبه احتمال این است که یک ژن خاص با چه احتمالی به هر کدام از خوشه‌ها تعلق دارد. این احتمال با استفاده از رابطه (۱۶) بدست می‌آید.
مرحله ماکسیمم‌سازی: در مرحله ماکسیمم‌سازی، برآوردهای $\hat{S}_g^{(r+1)}(t)$ و $\hat{V}_g^{(r+1)}$ و $\hat{\alpha}_g^{(r+1)}$ با برازش مدل (۱۱) به‌هنگام می‌شوند. تعداد تکرارها بین مرحله امیدگیری و ماکسیمم‌سازی تا جایی ادامه پیدا می‌کند که مقدار تابع درستنمایی همگرا شود. با توجه به این‌که الگوریتم EM برای داده‌های با بعد بالا ممکن است استوار نباشد، لذا برای یافتن پارامترهای مدل آمیخته متناهی از الگوریتم جدید EM رد-کنترل‌شده [۶] (RCEM) استفاده می‌نماییم که نسبت به الگوریم EM از استواری و سرعت محاسباتی بالاتری برای داده‌های با بعد بالا برخوردار است.

۲.۴ الگوریتم RCEM

این الگوریتم نتیجه اصلاح شده الگوریتم EM است که برای اولین بار توسط لیو و همکاران (۱۹۹۸) معرفی شد و سپس ما و همکاران (۲۰۰۶) از آن برای داده‌های بیان ژن دوره‌ای زمانی استفاده کردند. این الگوریتم شامل محاسبه احتمال‌های پسین (وزن‌ها) $\hat{\tau}_{ig}$، برای هر ژن در هر خوشه است به گونه‌ای که برای هر خوشه، اعضای درون خوشه با احتمال پسین به صورت

$$\hat{\tau}_{ig}^{*} = \begin{cases} \hat{\tau}_{ig} & \hat{\tau}_{ig}^{*} \geqslant c \text{ اگر} \\ c & \frac{\hat{\tau}_{ig}}{c} \text{ با احتمال } \hat{\tau}_{ig} \leqslant c \text{ اگر} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

به دست می‌آیند و با استفاده از رابطه $\hat{\tau}_{ig}^{*\prime} = \frac{\hat{\tau}_{ig}^{*}}{\sum_{g=۱}^{G} \hat{\tau}_{ig}^{*}}$ نرمال می‌شوند. مرحله ماکسیمم‌سازی در این الگوریتم مانند الگوریتم EM با استفاده از $\hat{\tau}_{ig}^{*}$ به عنوان وزن، از طریق ماکسیمم‌سازی رابطه (۱۵) صورت می‌گیرد. توجه شود که به ازای $c = ۰$ الگوریتم RCEM همان EM خواهد شد و به ازای $c = ۱$ این الگوریتم تبدیل به الگوریتم EM مونت کارلو [۷] (MCEM) خواهد شد (ما و همکاران، ۲۰۰۶). مقادیر

[6] Rejection-Controlled Expectation Maximization
[7] Monte Carlo EM

بزرگ c همگرایی سریع‌تر را نتیجه می‌دهد اما منجر به پایین آمدن دقت می‌شود.

۳.۴ انتخاب تعداد خوشه‌ها

هر چند مدلی که دارای تعداد خوشه بزرگتری باشد، دارای مطلوبیت بیشتری است اما افزایش تعداد خوشه‌ها باعث بیش برازندگی به داده‌ها خواهد شد. یک ملاک برای انتخاب تعداد خوشه‌های مناسب، ملاک $BIC = -۲\log L + d\log(\sum_{i=۱}^{N_T} n_i)$ است که در آن $\log L$ در (۱۵) آمده است، $d = kG +$ $۴G - ۱$ تعداد پارامترهای مدل و $\sum_{i=۱}^{N_T} n_i$ تعداد کل مشاهدات در n_i زمان است. برای تعیین تعداد خوشه مناسب که مقدار BIC برای مقادیر مختلف G محاسبه می‌شود، سپس با انتخاب مدلی که کمترین مقدار BIC را دارد، جواب بهینه برای تعداد خوشه بدست می‌آید.

۵ مطالعه شبیه‌سازی

در این بخش در مطالعه‌ای شبیه‌سازی به مقایسه عملکرد خوشه‌بندی با مدل‌های آمیخته ناپارامتری با استفاده از اسپلاین تاوانیده و مدل آمیخته نیمه‌پارامتری اسپلاین تاوانیده براساس معیار BIC می‌پردازیم. همچنین مدت زمان مورد نیاز برای مدل‌بندی و خوشه‌بندی هر یک از روش‌ها بیان خواهد شد.

برای بررسی عملکرد مدل پیشنهادی اثرات آمیخته نیمه‌پارامتری در خوشه‌بندی مطالعات طولی، ۱۰۰ مرتبه ۴۳۰ آزمودنی برای $i = ۱, \ldots, ۴۳۰$، در $n_i = n = ۱۰$ نقطه زمانی با فاصله یکسان در بازه $[۰, ۱]$ با استفاده از ۷ مدل آمیخته نیمه‌پارامتری تولید شده است به‌طوری که تعداد مشاهدات تولید شده برای هر یک از ۷ مدل $N_۱ = ۹۰$، $N_۲ = ۵۰$، $N_۳ = ۱۰۰$، $N_۴ = ۲۵$، $N_۵ = ۶۰$، $N_۶ = ۳۵$، $N_۷ = ۷۰$ است. ۷ مدل مورد نظر به صورت

$$y_{ij}^{(۱)} = x_{ij۰} + x_{ij۱} + x_{ij۲} + \sqrt{۲}\sin(۴\pi t_{ij}) + b_i + \varepsilon_{ij} \quad -۱$$

$$y_{ij}^{(۲)} = x_{ij۰} + ۲x_{ij۱} + x_{ij۲} + \tan(۲\pi t_{ij}) + b_i + \varepsilon_{ij} \quad -۲$$

$$y_{ij}^{(۳)} = x_{ij۰} + ۳x_{ij۱} + x_{ij۲} + b_i + \varepsilon_{ij} \quad -۳$$

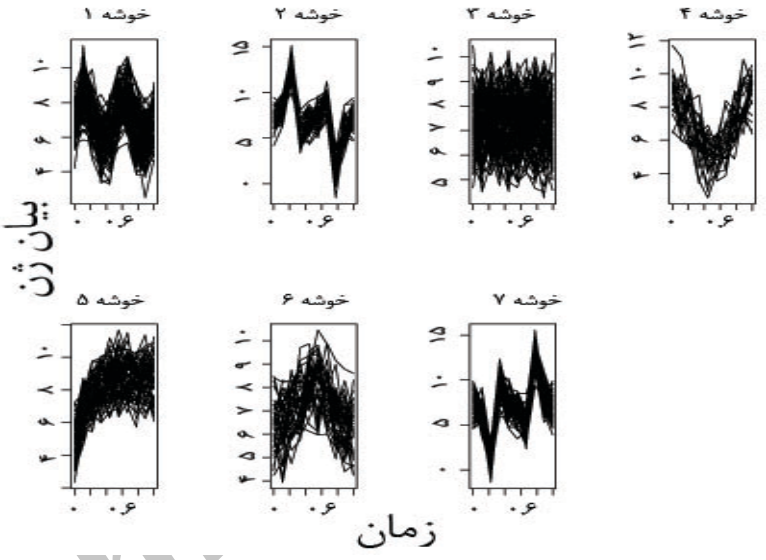
$$y_{ij}^{(۴)} = x_{ij۰} - x_{ij۱} + x_{ij۲} + ۲\cos(۲\pi t_{ij}) + b_i + \varepsilon_{ij} \quad -۴$$

$$y_{ij}^{(۵)} = x_{ij۰} - ۲x_{ij۱} + x_{ij۲} - ۴\exp(t_{ij}) + b_i + \varepsilon_{ij} \quad -۵$$

$$y_{ij}^{(۶)} = x_{ij۰} + x_{ij۱} + x_{ij۲} - ۲\cos(۲\pi t_{ij}) + b_i + \varepsilon_{ij} \quad -۶$$

$$y_{ij}^{(۷)} = ۲x_{ij۰} - x_{ij۱} + x_{ij۲} - \tan(۲\pi t_{ij}) + b_i + \varepsilon_{ij} \quad -۷$$

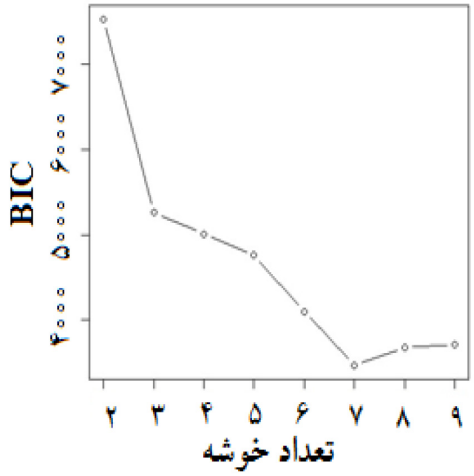
است که در آن متغیرهای کمکی دارای توزیع $x_{ij۰},\ x_{ij۱},\ x_{ij۲} \sim N(۰,۱)$ هستند. مقادیر پارامترهای مدل به‌گونه‌ای است که $b_i \sim N(۰,\sigma_b^۲)$ که در آن $\sigma_b = ۰/۶$ و $\varepsilon_{ij} \sim N(۰,\sigma_\varepsilon^۲)$ خطای اندازه‌گیری است که $\sigma_\varepsilon = ۰/۲$. نمودار هر یک از ۷ مدل معرفی شده در شکل ۱ ارائه شده‌اند. کدهای مربوط برای انجام خوشه‌بندی مدل اثرات آمیخته ناپارامتری و نیمه‌پارامتری در محیط برنامه‌نویسی R و با استفاده از بسته‌های نرم‌افزاری gamm۴ و Mvtnorm و Mclust نوشته شده و از نگارنده قابل دریافت هستند.



شکل ۱. نمودارهای داده‌های شبیه‌سازی شده برای ۷ مدل

پس از ترکیب کردن مشاهدات و تخصیص اولیه ژن‌ها به خوشه‌ها، مقادیر ابتدایی $\hat{S}_g^{(۰)}$، $\hat{\alpha}_g^{(۰)}$ و $\hat{\tau}_g^{(۰)}$ برآورد و مدل نیمه‌پارامتری اسپلاین تاوانیده (۱۱) به مشاهدات خوشه‌ها برازانده شد. سپس برای تمام خوشه‌ها $(g = ۲,\ldots,۹)$ بر اساس تابع چگالی $N(x_i\hat{\alpha}_g^{(۰)} + \hat{S}_g^{(۰)}(t_i),\ \hat{V}_{ig}^{(۰)})$ احتمالات پسین اولیه $(\hat{\tau}_{ig}^{(۰)})$ بدست آمد و با استفاده از الگوریتم RCEM و تعداد تکرارهای تعیین شده توسط کاربر (۱۵ تکرار) مراحل امیدگیری و ماکسیمم‌سازی انجام شده تا تابع درستنمایی به همگرایی برسد. سپس با استفاده از تابع درستنمایی، ماکسیمم‌ای بدست آمده است ومقادیر BIC برای تعداد خوشه‌های متفاوت به دست آمده است. همانطورکه در شکل ۲ مشاهده می‌شود مقادیر BIC در مقابل تعداد خوشه‌ها نشان‌دهنده قرار

گرفتن ۴۳۰ مشاهده در ۷ خوشه است زیرا این ۷ خوشه کم‌ترین مقدار BIC را دارد. نتایج نهایی داده‌های خوشه‌بندی شده به روش نیمه‌پارامتری در شکل ۳ ارائه شده است. در مدل نیمه‌پارامتری، تمام پارامترهای مربوط به متغیرهای کمکی معنی دار بوده است و براوردها به مقادیر واقعی نزدیک شده است.

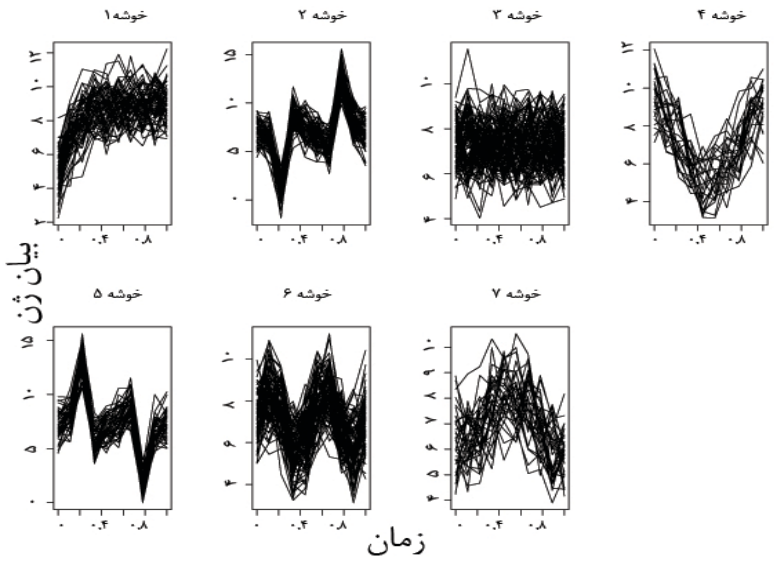


شکل ۲. نمودار BIC مدل نیمه‌پارامتری اسپلاین تاوانیده برای داده‌های شبیه‌سازی شده

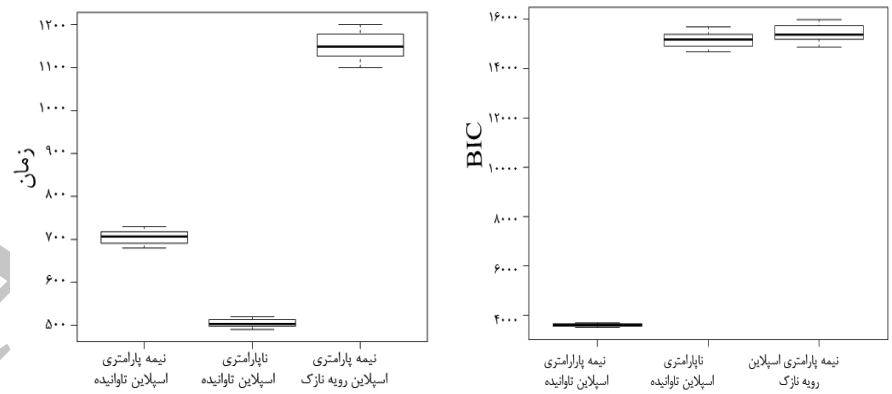
با برازش مدل ناپارامتری به داده‌ها (بدون در نظر گرفتن مولفه $X_g^T \alpha_g$ در رابطه (۱۱)) و مقایسه معیار BIC برای ۷ خوشه، مشخص گردید که کمترین مقدار BIC برای برازش مدل ناپارامتری به داده‌ها برابر ۱۴۶۷۶ است که در مقایسه با مقدار BIC برای مدل نیمه‌پارامتری اسپلاین تاوانیده عدد بزرگتری است. بنابراین مدل نیمه‌پارامتری به درستی به داده‌ها برازانده شده است.

برای مقایسه عملکرد مدل نیمه‌پارامتری با مدل‌های ناپارامتری اسپلاین تاوانیده با استفاده از هر یک از این ۳ مدل، ۴۳۰ داده‌ی تولید شده به روش نیمه‌پارامتری (که ۱۰۰ بار تولید کرده‌ایم) را خوشه‌بندی کرده‌ایم. نمودار جعبه‌ای مدت زمان تحلیل و مقادیر BIC هر یک از مدل‌ها برای ۱۰۰ مجموعه داده شبیه‌سازی شده به روش نیمه‌پارامتری در شکل ۴ آمده است.

همان‌طور که در جدول ۱ ملاحظه می‌شود مقدار BIC برای مدل نیمه‌پارامتری نسبت به ۲ مدل دیگر کوچکتر است، بنابراین نسبت به مدل‌های دیگر از کارایی بالاتری در تعیین ژن‌های مشابه در خوشه‌بندی برخوردار است.



شکل ۳. خوشه‌های حاصل از روش مدل‌بندی نیمه‌پارامتری برای داده‌های شبیه‌سازی شده



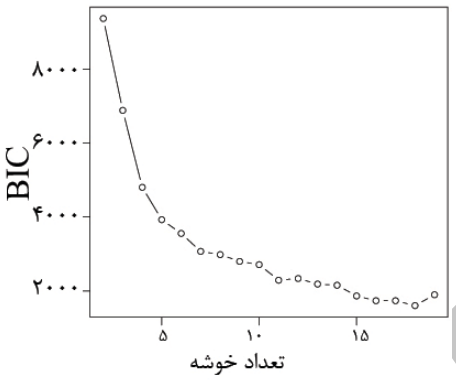
شکل ۴. نمودار جعبه‌ای زمان خوشه‌بندی و مقادیر BIC برای سه روش خوشه‌بندی

جدول ۱. مقادیر BIC و مدت زمان تحلیل داده‌های شبیه‌سازی شده نیمه‌پارامتری به روش‌های مختلف

روش	BIC	زمان(ثانیه)
ناپارامتری اسپلاین تاوانیده	۱۴۷۶۷	۵۰۵
ناپارامتری رویه‌ی نازک	۱۴۶۷۰	۱۱۴۶
نیمه‌پارامتری اسپلاین تاوانیده	۳۶۴۴	۷۱۲

۶ مثال کاربردی

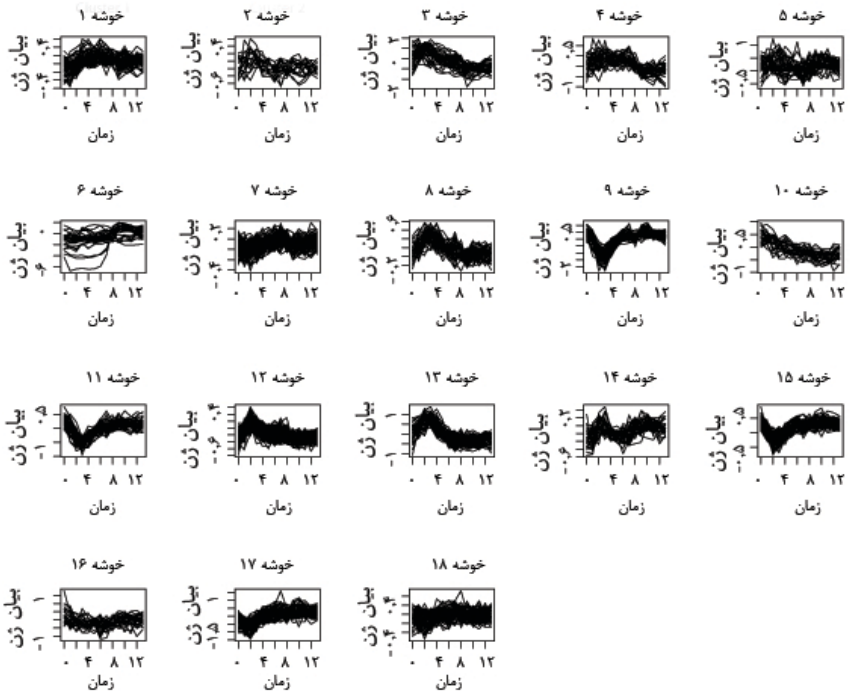
در این بخش اقدام به خوشه‌بندی داده‌های بیان ژن دوره‌ای زمانی با روش اثرات آمیخته‌ی نیمه‌پارامتری اسپلاین تاوانیده که دارای یک متغیر کمکی است، شده است. این داده‌ها مربوط به مقادیر بیان ژن یک مخمر در ۱۴ نسل در شرایط بی‌هوازی در محیط اسیدی است و هدف خوشه‌بندی، یافتن گروه‌های مشابه از ژن‌ها در طول زمان و خوشه‌بندی آن‌ها است. ژن‌های مربوط به این مخمر در نسل اول در شرایط هوازی بوده است که در نسل‌های بعدی محیط کشت از حالت هوازی به حالت O_2 بدون N_2 تغییر یافته است. محل کشت مخمر یک محیط اسیدی با میزان $5/7 < PH < 7$ در نظر گرفته شده است. این داده‌ها در پایگاه داده‌ای GEO با نام GSE۲۲۴۶ موجود است. به‌منظور بررسی اثر متغیر کمکی در مدل و عدم وجود متغیر کمکی در این مجموعه داده، مقدار PH توسط نویسندگان شبیه‌سازی و تولید شده است. لای و همکاران (۲۰۰۶)، ۲۳۸۸ ژن از این داده‌ها را که دارای بیان ژن متفاوت در طول زمان بوده‌اند با استفاده از روش خوشه‌بندی خودسازمانده گروه‌بندی کرده‌اند. آنها با استفاده از ملاک FCS، ۱۸ خوشه را بدست آورده‌اند. مطالعه آن‌ها بر روی ژن‌هایی بود که دارای بیان متفاوت در طول زمان هستند و ژن‌هایی که تغییرات کوچک‌تری را در طول زمان نشان می‌دادند را نادیده گرفتند. اما این ژن‌ها با توجه به اینکه رفتار مشابهی با ژن‌هایی که دارای بیان ژن متفاوت‌اند، دارند، ذاتاً مورد توجه زیست‌شناسان هستند. لذا در این مقاله با توجه به حجم بالای محاسبات و زمان‌بر بودن تحلیل این داده‌ها، ۱۰۰۰ ژن از ۵۶۱۷ ژن به تصادف انتخاب شده است و خوشه‌بندی بر روی این داده‌ها انجام شده است. با توجه به بررسی عملکرد روش‌های خوشه‌بندی در بخش ۵، داده‌های بیان ژن با استفاده از روش نیمه‌پارامتری اسپلاین تاوانیده خوشه‌بندی شده است. الگوریتم RCEM با ۵ نقطه شروع انجام شده است. همچنین $c = 0/5$، $k = 2$ و تعداد تکرارها در زمان انجام الگوریتم ۱۵ در نظر گرفته شده است. پس از بررسی نتایج برای خوشه‌های متفاوت، با استفاده از ملاک BIC، تعداد $G = 18$ براساس نتایج شکل ۵ انتخاب شده است. به طوری که ژن‌های درون هر خوشه دارای الگوهای بیانی مشابه در طول زمان هستند.



شکل ۵. نمودار BIC روش نیمه‌پارامتری برای داده‌های بیان ژن دوره‌ای زمانی مخمر

بحث و نتیجه‌گیری

از آنجا که خوشه‌بندی داده‌های بیان ژن بافت‌های سلولی به خوشه‌های همگن به منظور بررسی دامنه وسیعی از فرایندهای زیست‌شناختی، ضروری است، علاقه زیادی به خوشه‌بندی این داده‌ها برای به دست آوردن خوشه‌هایی با بیان ژن مشابه وجود دارد. در این مقاله روش جدیدی برای خوشه‌بندی مطالعات طولی با استفاده از مدل اثرات آمیخته‌ی نیمه‌پارامتری و اسپلاین تاوانیده برای داده‌های بیان ژن دوره‌ای زمانی، معرفی شده است که استفاده از اسپلاین تاوانیده باعث پایین آمدن زمان محاسبات خوشه‌بندی و استفاده از آمیخته‌ای از مدل اثرات تصادفی منجر به خوشه‌بندی و هموارسازی همزمان در داده‌ها می‌شود. از طرفی مدل اثرات آمیخته تصادفی منجر به ادغام تغییرات تصادفی مربوط به ژن‌های اطراف منحنی میانگین در هر خوشه می‌شود. از طرفی کارآیی مدل نیمه‌پارامتری نسبت به مدل ناپارامتری در داده‌هایی که دارای متغیر کمکی است و متغیر کمکی تاثیرگذار است، بالاتر است. یکی از مسائلی که در این گونه تحلیل‌ها می‌توان در نظر داشت خوشه‌بندی مطالعات طولی در حضور داده‌های دورافتاده و همچنین استفاده از روش‌های جانهی برای داده‌های گم‌شده است.



شکل ۶. نتایج خوشه‌های بدست آمده به روش نیمه‌پارامتری برای داده‌های بیان ژن دوره‌ای زمانی مخمر

تقدیر و تشکر

نویسندگان کمال تشکر و قدردانی را از داوران محترم، هیئت تحریریه و ویراستار مجله علوم آماری که سبب ارتقا سطح کیفی مقاله شدند، ابراز می‌دارند.

مراجع

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. and Staudt,

L. (2000), Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling. *Nature*, **403**, 503–511.

Bar-joseph, Z., Gerber, G., Gifford, D., Jaakkola, T., and Simon, I., (2003), Continuous Representation of Time Series Gene Expression Data, *Journal of Computational Biology*, **10**, 341-356

Chen, X. (2009), Curve-Based Clustering of Time Course Gene Expression Data Using Self Organizing Maps, *Journal of Bioinformatics and Computational Biology*, **7**, 645-661.

Coffey, N., Hinde, J., and Holian, E. (2014), Clustering Longitudinal Profiles UsingP-splines and Mixed Effects Models Applied to Time-Course Gene Expression Data, *Computational Statistics and Data Analysis*, **71**, 14-29.

Davidian, M. and Giltinan, D.M. (1995), *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998), Cluster Analysis and Display of Genome-wide Expression Patterns. *Proceeding of the National Academyof Sciences of the USA*. **95**, 14863–14868.

Eilers, P., and Marx, B. (1996), Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11**, 89–121.

Genolini, C., and Fallssard, B. (2010), KML: K-means for Longitudinal Data, *Computational Statistics*, **25**, 317-328.

Green, P., and Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models*. CRC Press, USA.

Harville, D. (1976), Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *Annual Statistics*, **4**, 384–395.

Harville, D.A. (1977), Maximum Likelihood Approaches to Variance Component Estimation and Related Problems. *Journal of the American Statistical Association*, **72**, 320–338.

Lai, L.C., Kosorukoff, A.L., Burke, P.V., and Kwast, K.E. (2006), Metabolic-state-dependent Remodeling of the Transcriptome in Response to Anoxia and Subsequent Reoxygenation in Saccharomyces Cerevisiae. *Eukaryot Cell*, **5**, 1468–1489.

Laird, N.M. and Ware, J.H. (1982), Random-effects Models for Longitudinal Data, *Biometrics*, **38**, 963-974.

Liu, J.S., Chen, R., and Wong, W.W. (1998), Rejection Control and Sequential Importance Sampling, *Journal of the American Statistical Association*, **93**, 1022-1031.

Ma, P.Castillo-Davis., Zhong, W., and Liu, J.S. (2006), A Data-Driven Clustering Method for Time Course Gene Expression Data, *Nucleic Acids Research*, **34**, 1261-1269.

Pfeifer, C. (2004), Classification of Longitudinal Profiles Based on Semi-Parametric Regression with Mixed Effects, *Statistical Modelling*, **4**, 314-323.

Ruppert, D. (2002), Selecting the Number of knots for Penalized Splines, *Journal of Computational and Graphical Statistics*, **11**, 735-757.

Ruppert, D., Wand, M., and Carroll, R. (2003), *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.

Shen, J. (2011), Additive Mixed Modeling of HIV Patient Outcomes Across Multiple Studies, *JRSSB*, **11**, 89-121.

Wahba, G. (1990), *Spline Models for Observational Data*. SIAM, Philadelphia.

Zhang, J.T. and Wu, H. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. John Wiley & Sons, New York.