

تشخیص داده های دورافتاده دایره ای با استفاده از یک مدل آمیزه ای از توزیع های فون میزس

خدیدجه عبدی، موسی گل علی زاده^۱ و تابان باغفلکی

گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۵/۱۱/۶ تاریخ پذیرش: ۱۳۹۷/۲/۵

چکیده: داده های دایره ای نوعی از داده های جهتی با دوره تناوبی مشخص هستند. به دلیل اینکه وجود داده های دور افتاده استنباط های آماری راجع به پارامترهای مدل های رگرسیون دایره ای را نامعتبر خواهد کرد، بررسی وجود آنها در تحلیل این مدل ها نیازمند توجه ویژه ای است. روش های متنوعی برای مدل بندی ساختار مجموعه داده ها شامل مشاهدات دور افتاده وجود دارد که به کارگیری مدل آمیزه ای یکی از مهم ترین آنهاست. در این مقاله علاوه بر مطالعه یکی از روش های مدل های رگرسیون دایره ای، به عنوان ایده پژوهشی جدید نحوه به کارگیری مدل آمیزه ای از توزیع های فون میزس برای تعیین داده های دورافتاده در این مدل تشریح می شود. برای برآورد پارامترهای مدل مذکور از الگوریتم *EM* استفاده خواهد شد. عملکرد مدل پیشنهادی با استفاده از مطالعات شبیه سازی بر روی و سپس از ماتریس درهم ریختگی برای ارزیابی اعتبار برازش مدل کمک گرفته می شود. به علاوه، روش های پیشنهادی در تحلیل داده های جهت موج سواحل استان بوشهر مورد استفاده قرار می گیرد.

واژه های کلیدی: داده های جهتی، نقاط دور افتاده، توزیع فون میزس، مدل های آمیزه ای، الگوریتم *EM*.

رده بندی ریاضی (۲۰۱۰): ۶۲H۱۱، ۶۲J۲۰.

۱- مقدمه

یکی از موضوعات مهم علمی مورد توجه محققین علوم زمین شناسی، زمین شناسی، هواشناسی و فیزیک، مطالعه بر روی داده های جهتی است. جهت های دو بعدی که به صورت زوایا نشان داده می شوند، با توجه به جهت صفر یعنی نقطه شروع و جهت دوران یعنی موافق یا

مخالف بودن آنها با حرکت عقربه‌های ساعت اندازه‌گیری می‌شوند. از آنجاکه جهت‌ها فاقد مقدار هستند، می‌توان آنها را به‌صورت نقاطی بر روی محیط دایره‌ای با شعاع واحد به مرکز مبدأ نمایش داد. به همین دلیل است که مشاهدات در فضای دو بعدی با طول واحد را داده‌های دایره‌ای^۱ می‌نامند. به دلیل ماهیت تناوبی بودن داده‌های دایره‌ای، باید تحلیل این نوع داده‌ها با روش‌هایی متفاوت از روش‌های معمول برای مجموعه داده‌های غیر دایره‌ای انجام شود. چنین تحلیلی در شاخه جدیدی از آمار با عنوان آمار دایره‌ای امکان‌پذیر است. برای تفکیک داده‌هایی که از ویژگی تناوبی بودن در فضاهای ناقلیدسی برخوردارند، بعد از فعالیت [۱] مر سوم شده است که اصطلاحات خاصی متداول شود. به‌طور دقیق‌تر، آمار مرتبط با تحلیل داده‌ها در فضاهای برداری را آمار خطی و برر سی مشاهداتی در فضاهای ناقلیدسی مانند دایره و کره را آمار غیرخطی می‌گویند [۲].

بحث راجع به مدل‌های رگرسیونی با متغیر پاسخ دایره‌ای اولین بار توسط [۳] برای تحلیل داده‌های دایره‌ای در حوزه علوم پزشکی مطرح شد. او توزیع فون‌میزس که معادل توزیع نرمال در فضای اقلیدسی است را برای تحلیل خطاها در فضای ناقلیدسی دایره معرفی کرد. [۱] مدل (۳) را به ازای مجموعه‌ای از مشاهده‌های وابسته دایره‌ای تعمیم داد. در ادامه این فعالیت‌ها [۴]، [۵]، [۶] و [۷] نیز مدل‌های رگرسیونی متفاوتی ارائه کردند. سپس، مطالعه جامع‌تری از ارتباط بین دو متغیر دایره‌ای توسط [۸] انجام شد. [۹] و [۱۰] این مدل را تعمیم داده و برآورد پارامترها را به روش ماکسیمم درست‌نمایی محاسبه کردند.

اگرچه مدل‌های رگرسیون دایره‌ای ارائه شده توانایی برآزش به برخی داده‌های جهت‌ی را دارند، اما وجود داده‌های دور افتاده از کارایی این مدل‌ها خواهد کاست. تشخیص داده‌های دور افتاده در مدل‌های رگرسیون دایره‌ای سابقه‌ای کم در آمار غیرخطی دارد. [۱۱] از نخستین کسانی بود که به برر سی چنین داده‌هایی در آمار دایره‌ای پرداخت و با معرفی چند آزمون آماری، راهکاری برای تشخیص داده‌های دور افتاده ارائه کرد.

[۱۲] با پیروی از رویکردی بیزی به تشخیص این نوع داده‌ها در آمار جهت‌ی پرداختند. [۲] نیز رهیافت نوینی برای رویارویی با چنین داده‌هایی ارائه کردند. در ادامه چنین فعالیت‌هایی، [۱۳] نیز آزمونی دیگر پیشنهاد کردند و [۱۴] به مطالعه و تشخیص داده دور افتاده در مدل رگرسیونی [۹] بر اساس آماره نسبت دترمینان کوواریانس‌ها پرداختند. تعدیل آماره‌ای در آمار خطی، به‌منظور تشخیص داده دور افتاده در مدل رگرسیون دایره‌ای [۵] توسط [۱۵] پیشنهاد شد. در رویکردی دیگر، [۱۶] آماره میانگین خطای دایره‌ای را برای تشخیص این نوع داده در مدل رگرسیون دایره‌ای ساده معرفی کردند. در راستای این تحقیقات مقاله حاضر مدل آمیزه‌ای را در

نظر می‌گیرد که توزیع فون میزس نقش اساسی را در آن بازی می‌کند. به‌طور دقیق‌تر، فرض می‌شود که داده‌های دایره‌ای از مدل آمیزه‌ای از توزیع‌های فون میزس با درصد آمیختگی نامعلوم آمده‌اند طوری که پارامترهای توزیع نیز مجهول هستند. آنگاه، درصد آمیختگی و پارامترهای هر یک از توزیع‌ها با استفاده از الگوریتم EM به صورت توأم برآورد می‌شوند. برای ارائه نتایج تحقیق، مقاله پیش رو به چهار بخش تقسیم شده است. بعد از مقدمه، در بخش دوم مدل رگرسیون دایره‌ای بر اساس بسط چند جمله‌ای مثلثاتی می‌آید. روش پیشنهادی این مقاله برای تشخیص داده دور افتاده در بخش سوم تشریح می‌شود. مطالعه شبیه‌سازی همراه با تحلیل یک مثال واقعی برای ارزیابی مطالب نظری ارائه شده در مقاله در بخش چهارم خواهد آمد. مقاله با نتیجه‌گیری کلی خاتمه می‌یابد.

۲- مدل رگرسیون دایره‌ای بر اساس بسط چند جمله‌ای مثلثاتی

کلیت [۵] برای دو متغیر تصادفی دایره‌ای x و y یک مدل رگرسیونی به‌صورت امید ریاضی e^{iy} به شرط x در نظر گرفتند، که در آن x متغیر تبیینی و y متغیر پاسخ بوده و $0 \leq x, y < 2\pi$. این مدل، که آن را در این مقاله به‌اختصار مدل رگرسیون دایره‌ای JS می‌نامیم، به‌صورت

$$E(e^{iy} | x) = \rho(x)e^{i\mu(x)} = g_1(x) + ig_2(x) \quad (1)$$

نوشته می‌شود که $\mu(x)$ میانگین جهتی y به شرط x و $0 \leq \rho(x) < 1$ پارامتر تمرکز شرطی نسبت به این جهت است. به‌علاوه،

$$e^{iy} = \cos y + i \sin y. \quad (2)$$

با اخذ امید ریاضی شرطی (به شرط x) از طرفین رابطه (۲) و مدنظر قرار دادن عبارت (۱)، به دست می‌آوریم:

$$E(\cos y | x) = g_1(x), \quad E(\sin y | x) = g_2(x), \quad (3)$$

که با توجه به روابط مثلثاتی می‌توان نوشت:

$$\mu(x) = \begin{cases} \tan^{-1} \frac{g_2(x)}{g_1(x)} & g_1(x) > 0, \\ \tan^{-1} \frac{g_2(x)}{g_1(x)} + \pi & g_1(x) \leq 0 \end{cases}$$

با توجه به دشواری برآورد ناپارامتری $g_1(x)$ و $g_2(x)$ در (۳) معمولاً از تقریب چند جمله‌ای مثلثاتی استفاده می‌شود. مناسبت این تقریب برای برآورد، یادآوری این حقیقت است که هر دو تابع $g_1(x)$ و $g_2(x)$ متناوب با دوره تناوب 2π هستند. با پیروی از [۵] و استفاده از چند جمله‌ای مثلثاتی می‌توان نوشت:

$$g_1(x) \approx \sum_{k=0}^m (A_k \cos kx + B_k \sin kx)$$

$$g_2(x) \approx \sum_{k=0}^m (C_k \cos kx + D_k \sin kx)$$

در نتیجه، با توجه به تساوی‌های ارائه شده در (۳) می‌توان فرض کرد:

$$\cos y = \sum_{k=0}^m (A_k \cos kx + B_k \sin kx) + \varepsilon_1 \quad (۴)$$

$$\sin y = \sum_{k=0}^m (C_k \cos kx + D_k \sin kx) + \varepsilon_2,$$

که در آن $\varepsilon = (\varepsilon_1, \varepsilon_2)^T$ بردار خطاهای تصادفی است که دارای توزیع نرمال با بردار میانگین $(0, 0)^T$ و ماتریس پراکندگی نامعلوم Σ است. یک مسئله مهم در برازش مدل رگرسیون چند جمله‌ای مثلثاتی تعیین درجه m است. برای این منظور، آزمون‌ها و معیارهای مربوطه باید طراحی شوند که قابلیت بهنگام سازی برآوردها را با افزایش درجه چند جمله‌ای داشته باشند. در رگرسیون چند جمله‌ای معمولی برای متغیرهای خطی از چند جمله‌ایهای متعامد برای این امر استفاده می‌شود. وقتی متغیرها دایره‌ای هستند، روشی که بهترین درجه m را برای بهنگام سازی برآوردها بدون محاسبه تمام ضرایب جدید به دست آورد، بررسی کاهش مجموع مربعات خطا به ازای افزایش m است. برای جزئیات بیشتر به [۵] مراجعه شود.

به دلیل اینکه وجود داده‌های دور افتاده منجر به استنباط‌های آماری نامعتبری راجع به پارامترها می‌شود، مطالعه آن‌ها مورد توجه محققین بی‌شماری بوده است. وجود چنین داده‌هایی در تحلیل آمار دایره‌ای نیازمند توجه بیشتری است چرا که مشاهدات دایره‌ای دارای ویژگی‌های خاصی مانند تناوبی بودن هستند. این در حالی است که داده‌های موجود در فضاهای برداری با این مشکل روبرو نیستند. روش‌های متنوعی برای مدل‌بندی ساختار مجموعه داده‌های دایره‌ای که شامل مشاهدات دور افتاده هستند وجود دارد. استفاده از توزیع‌های فون‌میزس آمیزه‌ای در داده‌های دایره‌ای به‌عنوان یک رویکرد جدید در این مقاله تشریح می‌شود.

۳- تشخیص داده دور افتاده بر اساس مدل آمیزه‌ای از توزیع‌های فون میزس

فرض کنید داده‌های در دسترس در خوشه‌هایی مجزا قرار بگیرند. آنگاه داده‌های موجود در خوشه‌ای با حجم بسیار پایین را می‌توان به‌عنوان داده‌های دور افتاده در نظر گرفت. در اینجا فرض می‌شود که متغیر تصادفی دایره‌ای θ از آمیزه دو توزیع فون میزس تولید شده باشد. در نتیجه، تابع چگالی احتمال آن با فرض اینکه $(\alpha_1, \alpha_2, p)^T$ که در آن $\alpha_h = (\mu_h, \kappa_h)^T$ ، $h = 1, 2$ بردار پارامترهای مدل باشد، به‌صورت

$$pf_1(\theta; \alpha_1^{(c)}) + (1-p)f_2(\theta; \alpha_2^{(c)}) \quad 0 < p < 1, 0 \leq \theta < 2\pi \quad (5)$$

خواهد بود که در آن $f_h(\theta; \alpha_h^{(c)})$ چگالی فون میزس با پارامترهای $0 \leq \mu_h < 2\pi$ و $\kappa_h \geq 0$ یعنی

$$f_h(\theta; \alpha_h^{(c)}) = \frac{1}{2\pi I_0(\kappa_h)} e^{\kappa_h \cos(\theta - \mu_h)}$$

است. در اینجا μ_h پارامتر میانگین، $\kappa_h \geq 0$ پارامتر تمرکز و $I_0(\kappa)$ تابع بسل تعدیل شده نوع اول، از مرتبه صفر است. برای جزئیات بیشتر [۱۷] را ببینید. برای برآورد پارامترهای این مدل آمیزه‌ای، از الگوریتم EM استفاده می‌شود [۱۸]. انتظار می‌رود این الگوریتم تفکیک داده‌ها در خوشه‌های مجزا و تشخیص خوشه‌ای با داده‌های دورافتاده را به‌طور مناسبی انجام دهد. با پیروی از این رویکرد، معمولاً داده‌های داخل خوشه‌ای که شامل حداکثر ۵ درصد از کل داده‌ها باشد به‌عنوان داده‌های دورافتاده در نظر گرفته می‌شوند. در ادامه جزئیاتی از نحوه به‌کارگیری الگوریتم EM برای مدل (۵) به قصد برآورد پارامترهای آن می‌آید.

فرض کنید $\theta_1, \dots, \theta_n$ دارای توزیع به فرم (۵) باشند و به ازای هر یک از θ_i متغیر پنهانی به‌صورت z_i موجود باشد طوری که

$$z_i = \begin{cases} 1 & \theta_i \sim f_1(\theta; \alpha_1^{(c)}) \\ 0 & \theta_i \sim f_2(\theta; \alpha_2^{(c)}) \end{cases}$$

بنا به ادبیات الگوریتم EM به مجموعه داده‌های مشاهده شده و متغیر پنهان (اصطلاحاً گمشده) داده‌های کامل گفته می‌شود. نمادگذاری $\underline{\theta} = (\theta_1, \dots, \theta_n)^T$ و $\underline{z} = (z_1, \dots, z_n)^T$ را در نظر بگیرید. آنگاه، لگاریتم تابع درستنمایی داده کامل به‌صورت

$$\ln(\underline{\theta}, \underline{z} | \alpha, p) = \sum_{i=1}^n z_i [\log p + f_1(\theta_i, \alpha_1)] + (1-z_i) [\log(1-p) + \log f_2(\theta_i, \alpha_2)] \quad (6)$$

است. با پیروی از [۱۸]، الگوریتم EM شامل دو قدم معروف گام E یا امیدگیری و گام M با بیشینه‌سازی است. برای تشریح جزئیات این الگوریتم، در ادامه مراحل محاسبه این دو گام برای تشخیص داده‌های دور افتاده دایره‌ای تشریح می‌شوند.

گام E : محاسبه امید ریاضی رابطه (۶) یعنی $Q(\alpha, \alpha^{(c)}) = E(\ln(\underline{\theta}, \underline{z} | \alpha, p))$ توسط تساوی

$$Q(\alpha, \alpha^{(c)}) = \sum_{i=1}^n t_{i_1}^{(c)} ((\log p + \log f_{\nu}(\theta_i, \alpha)) + t_{i_2}^{(c)} (\log p(1-p) + \log f_{\nu}(\theta_i, \alpha)), \quad (7)$$

صورت می‌گیرد که در این عبارت $t_{ih}^{(c)}$ احتمال شرطی $z_i = 1$ به شرط داشتن مقدار اولیه پارامترهای $\alpha^{(c)}$ و مشاهدات θ است. با توجه به محاسبات احتمالات پسین می‌توان نوشت:

$$t_{ih}^{(c)} = p(z_{ih} = 1 | \theta, \alpha^{(c)}) = \frac{p_h^{(c)} f_h(\theta_i; \alpha_h^{(c)})}{\sum_{h=1}^k p_h^{(c)} f_h(\theta_i; \alpha_h^{(c)})}$$

گام M : در این گام ماکسیمم‌سازی رابطه (۷) بر حسب پارامترها مدنظر است. به زبانی دقیق‌تر، هدف محاسبه عبارت $\alpha^{(c+1)} = \arg \max(Q(\alpha, \alpha^{(c)}))$ است. در زیر نشان داده می‌شود که نتیجه این امر برای هر یک از پارامترها دارای فرم بسته است.

با مشتق‌گیری از رابطه (۷) نسبت به پارامتر μ_h به ازای $h = 1, 2$ داریم:

$$\begin{aligned} \frac{\partial Q(\alpha, \alpha^{(c)})}{\partial \mu_h} &= \sum_{i=1}^n t_{ih}^{(c)} \left\{ \frac{-\kappa_h (-\sin(\theta_i - \mu_h)) e^{\kappa_h \cos(\theta_i - \mu_h)}}{e^{\kappa_h \cos(\theta_i - \mu_h)}} \right\} \\ &= \sum_{i=1}^n t_{ih}^{(c)} \{ \kappa_h \sin(\theta_i - \mu_h) \}. \end{aligned}$$

با مساوی صفر قرار دادن رابطه اخیر، یعنی

$$\sum_{i=1}^n t_{ih}^{(c)} \{ \kappa_h \sin(\theta_i \cos \mu_h - \cos \theta_i \sin \mu_h) \} = 0$$

پارامتر μ به صورت

$$\hat{\mu}_h^{(c+1)} = \tan^{-1} \left(\frac{\sum_{i=1}^n t_{ih}^{(c)} \sin \theta_i}{\sum_{i=1}^n t_{ih}^{(c)} \cos \theta_i} \right) = \tan^{-1} \left(\frac{S}{C} \right)$$

به‌روز خواهد شد. با توجه به اینکه $\hat{\mu}_h^{(c+1)}$ میانگین جهتی است، با پیروی از [۱۷]، باید برآورد آن به‌صورت

$$\hat{\mu}_h^{(c+1)} = \begin{cases} \tan^{-1} \left(\frac{S}{C} \right) & S > 0, C > 0, \\ \tan^{-1} \left(\frac{S}{C} \right) + \pi & C < 0 \\ \tan^{-1} \left(\frac{S}{C} \right) + 2\pi & S < 0, C > 0, \end{cases}$$

بازنویسی شود، که در آن $S = \sum_{i=1}^n t_{ih}^{(c)} \sin \theta_i$ و $C = \sum_{i=1}^n t_{ih}^{(c)} \cos \theta_i$.

برای به‌روزرسانی پارامتر κ_h نیز باید از رابطه (۷) نسبت به این پارامتر مشتق گرفت. می‌توان ملاحظه کرد که

$$\frac{\partial Q(\alpha, \alpha^{(c)})}{\partial \kappa_h} = \sum_{i=1}^n t_{ih}^{(c)} \left\{ \frac{e^{\kappa_h \cos(\theta_i - \mu_h)} \left(\frac{-2\pi I_1(\kappa_h)}{(2\pi I_0(\kappa_h))^2} + \frac{\cos(\theta_i - \mu_h)}{2\pi I_0(\kappa_h)} \right)}{\frac{e^{\kappa_h \cos(\theta_i - \mu_h)}}{2\pi I_0(\kappa_h)}} \right\} \quad (8)$$

$$= \sum_{i=1}^n t_{ih}^{(c)} (\cos(\theta_i - \mu_h) - A(\kappa_h)).$$

که به‌طور کلی $A_p(x) = \frac{I_p(x)}{I_0(x)}$ که در اینجا اندیس یک برای تابع A حذف می‌شود. با مساوی صفر قرار دادن عبارت (۸)، تساوی

$$A(\hat{\kappa}_h) \sum_{i=1}^n t_{ih}^{(c)} = \sum_{i=1}^n t_{ih}^{(c)} \cos(\theta_i - \hat{\mu}_h)$$

حاصل می‌شود که از طریق آن پارامتر κ_h به‌صورت

$$\hat{K}_h^{(c+1)} = A^{-1} \left(\frac{\sum_{i=1}^n t_{ih}^{(c)} \cos(\theta_i - \hat{\mu}_h)}{\sum_{i=1}^n t_{ih}^{(c)}} \right),$$

به‌روز خواهد شد. پارامتر آمیختگی نیز مشابه دو پارامتر تمرکز و میانگین جهتی با مشتق‌گیری از رابطه (۷) به دست می‌آید. نتیجه حاصل به‌صورت زیر است.

$$\hat{P}_h^{(c+1)} = \frac{\sum_{i=1}^n t_{ih}^{(c)}}{\sum_{i=1}^n t_{i\lambda}^{(c)} + t_{i\gamma}^{(c)}} = \frac{1}{n} \sum_{i=1}^n t_{ih}^{(c)}.$$

اکنون می‌توان تشخیص داده دور افتاده با استفاده از روش الگوریتم EM را به‌صورت زیر تدوین کرد. به‌ازای تعداد نمونه تولید شده یعنی $i = 1, \dots, n$ و دو خوشه $h = 1, 2$ احتمالات پسین $t_{ih}^{(c)}$ محاسبه شود. سپس لازم است ماکسیمم مقادیر این احتمالات در هر مرحله محاسبه و مقادیر حاصل در یک خوشه قرار داده شوند. با انجام این کار، درنهایت دو خوشه به دست می‌آید. خوشه‌ای که داده کمتری دارد به‌عنوان خوشه دور افتاده شناخته می‌شود و داده‌های آن خوشه، داده‌های دور افتاده خواهند بود.

۴- مطالعه شبیه‌سازی و تحلیل مثال واقعی

در این بخش با انجام مطالعات شبیه‌سازی به بررسی روش تشخیص داده دور افتاده مبتنی بر مدل آمیزه‌ای از توزیع فون‌میزس با روش الگوریتم EM پرداخته می‌شود. صحت تشخیص داده دورافتاده با استفاده از این روش پیشنهادی با ماتریس درهم ریختگی بررسی می‌شود. علاوه بر این، تحلیل یک مثال واقعی مدنظر است. داده‌های این مثال مربوط به جهت موج و باد ثبت شده توسط بویه بوشهر است که از سازمان هواشناسی کشور تهیه شد. مدل رگرسیون دایره‌ای بر این داده‌ها برازش داده می‌شود و با استفاده از مانده‌های رگرسیونی، داده دور افتاده تشخیص داده می‌شود. تحلیل عملکرد روش نیز مورد بحث قرار می‌گیرد.

۴-۱- مطالعه شبیه‌سازی

برای انجام مطالعه شبیه‌سازی فرض می‌کنیم مجموعه داده‌ها از آمیزه دو توزیع فون‌میزس تولید شده‌اند. به‌طور دقیق‌تر، در نظر می‌گیریم داده‌ها از طریق رابطه

چنین فرض کرد که ۵ درصد از داده‌های تولیدی در یک خوشه و ۹۵ درصد از آنها در خوشه دیگر قرار دارند. لذا، مشاهداتی که در خوشه ۵ درصد هستند را به‌عنوان داده‌های دورافتاده در نظر گرفته‌ایم. برای تشخیص داده دورافتاده بر اساس مدل آمیزه‌ای از توزیع‌های فون میزس از روش الگوریتم EM که در بخش قبل تشریح شد، استفاده می‌کنیم. برای شروع شبیه‌سازی این مدل آمیزه، مقادیر اولیه زیر در نظر گرفته شد:

$$\mu_1 = 0/5\pi, \kappa_1 = 14/5, \mu_2 = 1/0.1\pi, \kappa_2 = 116, p_1 = 0/45$$

با این مقادیر اولیه، برآورد پارامترها به روش الگوریتم EM به‌طور متوسط در ۳۰ تکرار همگرا شدند. برآورد پارامترها به ازای حجم نمونه متفاوت و ۵۰۰ بار تکرار شبیه‌سازی در جدول ۱ آمده است. توجه شود که مقادیر داخل پرانتز، میانگین مربعات خطا (MSE) را نشان می‌دهد. همان‌طور که ملاحظه می‌شود با افزایش حجم نمونه، برآورد پارامترها با MSE کمتری به مقدار واقعی نزدیک می‌شوند. واضح است که چون خوشه با حجم ۰/۵ نمونه، تعداد کمتری دارد، برآورد پارامتر κ_1 نسبت به برآورد κ_2 با سرعت کمتری به مقدار واقعی نزدیک شده است. لذا، انتظار می‌رود با افزایش حجم نمونه برآورد پارامتر اولی با مقدار MSE کمتری مقدار واقعی را برآورد کند. همچنین، به علت اینکه برآورد نسبت آمیختگی (p) با MSE بسیار پایینی برآورد شده است، انتظار می‌رود که الگوریتم EM با خطای ناچیزی بتواند داده دورافتاده را تشخیص دهد.

جدول (۱): برآورد پارامترهای مدل آمیزه فون میزس بر اساس الگوریتم EM ، مقادیر داخل پرانتز مقدار MSE را نشان می‌دهد.

پارامتر	مقدار واقعی	$n = 100$	$n = 500$	$n = 1000$
μ_1	۱/۵۷	۱/۶(۰/۰۴)	۱/۵۷(۰/۰۰۶)	۱/۵۷(۰/۰۰۲۴)
κ_1	۱۰/۳	۱۸/۵۶(۱۲۷/۵۲)	۱۱/۸۶(۱۹/۷)	۱۰/۹۶(۷/۲۵)
μ_2	۳/۱۴	۳/۱۳(۰/۰۰۱)	۳/۱۴(۰/۰۰۰۱۷)	۳/۱۴(۰/۰۰۰۰۹)
κ_2	۱۲/۸	۱۳/۱۲(۵/۲۱)	۱۲/۵۶(۰/۷۱)	۱۲/۶۴(۰/۳۶)
p_1	۰/۰۵	۰/۰۶(۰/۰۰۸)	۰/۰۵(۰/۰۰۰۱)	۰/۰۵(۰/۰۰۰۰۴)

برای اطمینان از تفکیک درست پارامترها، می‌توان از ماتریس درهم‌ریختگی استفاده کرد. با استفاده از ماتریس درهم‌ریختگی معیارهایی برای اعتبار سنجی تشخیص خوشه شامل داده‌های

دورافتاده برای حجم نمونه‌های متفاوت در هر تکرار شبیه سازی، محاسبه شده است. میانگین این معیارها به ازای ۵۰۰ بار تکرار در جدول ۲ آورده شده است. در این جدول، معیار درستی^۱ (دقت خوشه‌بندی داده‌ها)، حساسیت^۲ (نسبت تشخیص درست خوشه آلوده)، شیوع^۳ (نسبت وقوع خوشه آلوده) و نرخ مثبت آلوده^۴ (نسبت تشخیص غلط خوشه شامل داده دورافتاده) را با استفاده از روش الگوریتم *EM* نشان می‌دهند. با توجه به نتایج جدول ۲ ملاحظه می‌شود که معیار نرخ مثبت آلوده مقدار ۰/۰۰۱ را نشان می‌دهد، به این معنی که الگوریتم *EM* در ۰/۰۰۱ موارد داده‌های دورافتاده را تشخیص نداده است که این مقدار خطا بسیار ناچیز است.

جدول (۲): میانگین معیارهای اعتبارسنجی تشخیص خوشه مربوط به داده‌های دور افتاده

معیارهای اعتبارسنجی	$n = 100$	$n = 500$	$n = 1000$
درستی	۰/۹۹	۰/۹۹	۰/۹۹
حساسیت	۰/۹۶	۰/۹۷	۰/۹۸
شیوع	۰/۰۴۷	۰/۰۴۹	۰/۰۵
نرخ مثبت کاذب	۰/۰۰۲	۰/۰۰۱	۰/۰۰۱

معیار درستی مقدار دقت خوشه‌بندی داده‌ها را ۰/۹۹ محاسبه کرده است. این عدد نشان می‌دهد که روش پیشنهادی در این مقاله توانسته است داده‌های دورافتاده را تا حد زیادی به درستی تشخیص دهد.

۴-۲- تحلیل یک مثال واقعی

در این بخش داده‌های مربوط به جهت باد و جهت موج که با استفاده از بویه بوشهر و بر اساس درجه طی یک سال اندازه‌گیری شده مورد بررسی قرار می‌گیرد. در واقع این داده‌ها ماهیتاً داده دایره‌ای هستند. در واقع، هدف بررسی نوع ارتباط بین متغیرهای این داده‌ها است. داده‌های مثال مورد مطالعه، مربوط به جهت موج و باد است که توسط بویه بوشهر این داده‌ها در ۲۴ ساعت هر روز و در طول سال ۲۰۰۸ ثبت شدند. از نقطه نظر جغرافیایی، بویه بوشهر در طول جغرافیایی ۵۰/۴۹ و عرض جغرافیایی ۲۸/۵ قرار دارد.

۱ - Accuracy

۲ - Sensitivity

۳- Prevalence

۴- False positive rate

داده‌های دایره‌ای این مثال همگی برحسب درجه هستند. اما، برای سهولت محاسبات و هم‌چنین جلوگیری از نمادگذاری درجه، تمامی داده‌ها به رادیان تبدیل شدند. علاوه بر این، برای از بین بردن اثر روز تمامی داده‌های ثبت شده در ساعت $۱۶:۳۰$ در نظر گرفته شد. یکی از مفروضات اولیه رگرسیون دایره‌ای استقلال آماری بین مشاهدات است. از آنجائی که انتظار می‌رود داده‌های روزهای متوالی به طریقی همبستگی داشته باشند سعی شد همبستگی حداقل یک روزه داده‌ها بررسی شود. پی مقدار آزمون همبستگی پیرسون برای تأخیر یک برابر $۱۰^{-۱۲}$ تا $۱۰^{-۳}$ ، تأخیر دو برابر $۱۰^{-۳}$ تا $۱۰^{-۲}$ ، تأخیر سه برابر $۱۰^{-۳}$ تا $۱۰^{-۲}$ و تأخیر چهار برابر $۱۰^{-۲}$ به دست آمد. این مقادیر نشان می‌دهد که از تأخیر سه به بعد بین داده‌ها همبستگی وجود ندارد. به عبارتی دیگر، همبستگی داده‌ها بعد از گام چهارم از بین می‌رود. این نتیجه نیز در شکل ۱ مشاهده می‌شود. از آنجائی که تعداد کل داده‌ها در ساعت موردنظر ۳۶۰ عدد بود، با انتخاب هر چهارمین مشاهده تعداد داده‌ها به ۹۰ کاهش یافت. بعد از تحلیل مقدماتی برای پلایش داده‌ها یکی از علاقه‌مندی‌های محاسبه خلاصه‌های آمار دایره‌ای برای این داده‌ها است.

میانگین، میانه، دامنه و واریانس دایره‌ای به‌اضافه میانگین طول برآیند برای داده‌های جهت موج و جهت باد محاسبه و نتایج در جدول ۳ آورده شده است.

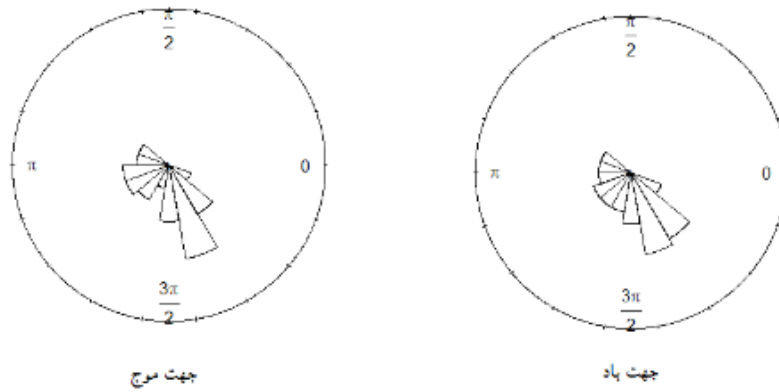
جدول (۳): تحلیل توصیفی داده‌های جهت باد و جهت موج (برحسب رادیان)

متغیر	میانگین	میانه	دامنه	واریانس	میانگین طول برآیند
جهت موج	-۱/۶۷	-۱/۳۷	۲/۹	۰/۳	۰/۷
جهت باد	-۱/۵	-۱/۲۶	۳/۱۷	۰/۲۹	۰/۷

در ستون اول و دوم معیارهای مرکزی این دو متغیر محاسبه شده است. با توجه به مقدار میانگین دایره‌ای، جهت باد و موج غالب منطقه ساحلی بوشهر در فصل به سمت جنوب بوده است. در ستون‌های بعدی جدول ۲ معیارهای پراکندگی آورده شده است. کوچک‌ترین کماتی که شامل همه مشاهدات نمونه‌ای شود، دامنه دایره‌ای است که مقادیر آن برای هر دو متغیر جهت موج و باد در دو فصل در ستون سوم آمده است. با توجه به اینکه واریانس دایره‌ای جهت موج فصل پاییز به یک نزدیک‌تر است پراکندگی بیشتری نسبت به جهت‌های دیگر را دارد.

نمودار گل‌سرخ^۱ یا در اصطلاح جغرافیا گلباد برای هر دو متغیر پاسخ و متغیر تبیینی در شکل ۱ نشان داده شده است. به‌وضوح می‌توان جهت غالب و پراکندگی برای متغیرهای فصل را در این شکل مشاهده کرد.

۱-Rose diagram



شکل (۱): نمودار گلباد جهت موج و باد فصل

برای این امر، با پیروی از [۷]، ابتدا آماره $AICC$ دو مدل رگرسیون دایره‌ای [۵] و [۹] باهم مقایسه می‌شوند. مدلی که مقدار

$$AICC = 2n \log I_0(\hat{\kappa}) - 2n\hat{\kappa} + \frac{n(n+1)}{n-l-2}$$

کمتری داشته باشد، برازش بهتری به داده‌ها ارائه می‌کند. در این آماره n حجم نمونه و l تعداد پارامترهای برآورد شده هستند. برای مدل [۵] این آماره، $-118/85$ و برای مدل [۹]، مقدارش $-113/48$ به دست آمدند. در نتیجه، باید مدل رگرسیون دایره‌ای JS به داده‌ها برازش داده شود. در برازش این مدل رگرسیونی دایره‌ای، جهت باد به‌عنوان متغیر تبیینی و جهت موج به‌عنوان متغیر پاسخ در نظر گرفته می‌شوند. پارامترهای مدل با فرض اینکه درجه چند جمله‌ای مثلثاتی برابر یک است ($m=1$) به‌صورت زیر برآورد شدند:

$$\begin{aligned} \hat{A}_0 &= -0/163, \hat{A}_1 = 0/852, \hat{B}_1 = -0/071, \\ \hat{C}_0 &= -0/270, \hat{C}_1 = -0/289, \hat{D}_1 = 0/581. \end{aligned}$$

بعد از برآورد پارامترها، توابع $g_1(x)$ و $g_2(x)$ به‌صورت

$$\begin{aligned} \hat{g}_1(x) &= -0/163 + 0/852 \cos x - 0/071 \sin x \\ \hat{g}_2(x) &= -0/270 - 0/289 \cos x - 0/581 \sin x \end{aligned}$$

برآورد شدند. با توجه به اینکه p -مقدار آزمون χ^2 (برای نحوه محاسبه این آزمون به [۵] مراجعه شود) با سه درجه آزادی برای $\hat{g}_1(x)$ و $\hat{g}_2(x)$ به ترتیب $0/23$ و $0/92$ است، نیازی

به افزایش درجه مثلثاتی از یک به دو نیست. بعد از برآورد $g(x)$ ها، مدل $[5]$ ، یعنی $E(e^{iy} | x) = \rho(x)e^{i\mu(x)} = g_1(x) + ig_2(x)$ محاسبه شد، که برای این منظور داریم:

$$\hat{\mu}(x) = \begin{cases} \tan^{-1} \left(\frac{-0.270 - 0.289 \cos x_j + 0.581 \sin x_j}{-0.163 + 0.852 \cos x_j - 0.071 \sin x_j} \right), & 0.89 < x_j < 3.17 \\ \pi + \tan^{-1} \left(\frac{-0.270 - 0.289 \cos x_j + 0.581 \sin x_j}{-0.163 + 0.852 \cos x_j - 0.071 \sin x_j} \right), & x_j \leq 0.89 \text{ or } 3.17 \leq x_j \leq 2\pi \end{cases}$$

9

$$\hat{\rho}(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{\rho}^2(x_i)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (g_1^2(x_i) + g_2^2(x_i))} = 0.94.$$

اکنون از رویکرد مدل آمیزه‌ای از توزیع‌های فون میزس برای تعیین داده‌های دورافتاده استفاده می‌کنیم. با فرض اینکه مانده‌های مدل رگرسیونی JS دارای دو خوشه هستند، با استفاده از الگوریتم EM به تشخیص داده‌های دور افتاده پرداختیم. با استفاده از این الگوریتم، ملاحظه شد که تنها مشاهده شماره ۲۵ در یک خوشه قرار گرفت و لذا می‌توان آن را به‌عنوان داده دور افتاده در نظر گرفت. این مشاهده مربوط به روز جمعه ۲۰ اردیبهشت سال ۱۳۸۷ است.

۵- بحث و نتیجه‌گیری

برخورد با داده‌های دور افتاده در یک تحلیل آماری از مسائلی بحث‌برانگیز در آمار است. این موضوع در داده‌های دایره‌ای از اهمیت دوچندان برخوردار است، چرا که ماهیت تناوبی بودن داده‌ها بر پیچیدگی موضوع می‌افزاید. در این مقاله رویکرد جدیدی برای برخورد با این مطلب پیشنهاد شد. یکی از مدل‌های رگرسیونی دایره‌ای مطالعه و استفاده از آن مدل برای تعیین نقاط دورافتاده دایره‌ای تشریح شد. برای ارزیابی عملکرد روش‌های پیشنهادی مطالعه شبیه‌سازی انجام گرفت. با توجه به نتایج شبیه‌سازی ملاحظه شد که تشخیص داده دورافتاده بر اساس مدل آمیزه‌ای از توزیع فون میزس با روش الگوریتم EM خطای کمی به همراه دارد.

تعمیم تعداد خوشه‌ها از دو به مقادیر بیشتر می‌تواند موضوع مناسبی برای تحقیقات آتی در این حوزه باشد. به‌علاوه، انتظار می‌رود استفاده از روش بی‌زی در مدل آمیزه فوق‌میزس، میزان خطای برآورد پارامترهای تمرکز خوشه‌ها را به‌اندازه کافی کاهش دهد. اما بررسی این موضوع نیاز به تحقیقات وسیع‌تری دارد. در بررسی مثال واقعی سعی شد همبستگی زمانی بین داده‌ها به طریقی از بین برود، تا بدین طریق فرض استقلال مشاهدات مصداق داشته باشد. پیشنهاد می‌شود برای مطالعات آینده و بررسی دقیق‌تر این داده‌ها، از تحلیل سری‌های زمانی دایره‌ای

استفاده شود. بحث راجع به تشخیص داده‌های دور افتاده در سری زمانی دایره‌ای نیز می‌تواند موضوع جالبی برای تحقیقات آتی باشد.

منابع

- [1] Mardia, K. V. (1972), *Statistics of Directional Data*, London: Academic Press.
- [2] Jammalamadaka, S. R. and SenGupta, A. (2001), *Topics in Circular Statistics*, Singapore: World Scientific.
- [3] Gould, A. L. (1969), A Regression Technique for Angular Variates, *Biometrics*, **25**, 683-700.
- [4] Laycock, P. J. (1975), Optimal Design: Regression Models for Directions, *Biometrika*, **62**, 305-311.
- [5] Sarma, Y., and Jammalamadaka, S. (1993), Circular Regression, *Statistical Science and Data Analysis*, 109-128.
- [6] Rivest, L. P. (1997), A Decentred Predictor for Circular-Circular Regression, *Biometrika*, **84**, 717-726.
- [7] Lund, U. (1999), Least Circular Distance Regression for Directional Data, *Journal of Applied Statistics*, **26**, 723-733.
- [8] Downs, T. D. and Mardia, K. V. (2002), Circular Regression, *Biometrika*, **89**, 683-697.
- [9] Hussain, A. G., Fieller, N. R. J. and Stillman, E. C. (2004), Linear Regression for Circular Variables with Application to Directional Data, *Journal of Applied Science and Technology*, **9**, 1-6.
- [10] Hussain, A. G., Abdullah, N. A. and Mohamed, I. (2010), A Complex Linear Regression Model, *Sains Malaysian*, **39**, 491-494.
- [11] Collett, D. (1980), Outliers in Circular Data, *Journal of Applied Statistics*, **29**, 50-57.
- [12] Bagchi, P. and Guttman, I. (1990), Spuriousity and Outliers in Directional Data, *Journal of Applied Statistics*, **17**, 341-350.
- [13] Abuzaid, A. H., Mohamed, I. B. and Hussain, A. G. (2009), A New Test of Discordancy in Circular Data, *Communications in Statistics-Simulation and Computation*, **38**, 682-691.
- [14] Abuzaid, A., Mohamed, I., Rambli, A. and Hussain, A. G. (2011), COVRATIO Statistic for Simple Circular Regression Model, *Chiang Mai Journal of Science*, **38**, 321-330.

- [15] Ibrahim, S., Rambli, A., Hussain, A. G. and Mohamed, I. (2013), Outlier Detection in a Circular Regression Model Using COVRATIO Statistic, *Communication in Statistics-Simulation and Computation*, 42, 2272-2280.
- [16] Abuzaid, A. H., Hussain, A. G. and Mohamed I. B. (2013), Detection of Outliers in Simple Circular Regression Models Using the Mean Circular Error Statistic, *Journal of Statistical Computation and Simulation*, 83, 269-277.
- [17] Mardia, K. V. and Jupp, P. (2000), *Directional Statistics*, New York: John Wiley.
- [18] McLachlan G. J. and Krishnan T. (2008) *The EM Algorithm and Extensions*, New York: John Wiley.

Circular Outliers Detection Using a Mixture of Von Mises Distributions

Khadigeh Abdi, Moussa Golarizadeh, and Taban Baghfalaki

Department of Statistics, Tarbiat Modares University, Tehran, Iran.

Abstract

Circular data are typical examples of directional data with specific systematic period. Because the presence of outliers leads to invalid statistical inferences on the parameters of the circular regression models, their prevalence in analyzing these models requires particular attentions. There are different approaches for modeling the structure of the data set with outliers in which using the mixture models is among the most important ones. As a new idea, to study the methods in determining the outlier data and to employ the mixture model of Von Mises distribution are considered in this paper. The EM algorithm is utilized for estimating the parameters of these models. The performance of the proposed models is investigated using some simulation studies and then the confusion matrix is considered to evaluate the accuracy of fitting models. Also, the proposed methods are applied to analysis a real data set, related to the wave directional analysis in the Bushehr province in Iran.

: **Keywords** Directional data; Outlier points; Von Mises distribution; Mixture models; *EM* algorithm.

Mathematics Subject Classification (2010): 62H11, 62J20.

کد الگوریتم *EM* برای برآورد پارامترهای مدل

```

iteration=500
M=matrix(1,nrow=iteration,ncol=5)
colnames(M)=c("mu1","mu2","kappa1","kappa2","p1")
for(kk in 1:iteration) {
n=10000
h=10000
k=2
meanobs1<-0.5*p1
meanobs2<-p1
kappaobs1<-10.3
kappaobs2<-12.8
pobs<-0.1
z<-rbinom(n,1,pobs)
randata1<-rvm(n,meanobs1,kappaobs1)
randata2<-rvm(n,meanobs2,kappaobs2)
theta<-(z*randata1+(1-z)*randata2)%%(2*pi)
mu=kappa=rep(0.2)
Results=matrix(1,h,5)
colnames(Results)=c("mu1","mu2","kappa1","kappa2","p1")
Results[1,]=c(0.35*pi,10.1*pi,10.3,10.1,0.1)
for(c in 2:h)
}
p=c(Results[(c-1),5],1-Results[(c-1),5])
mu=Results[(c-1),1:2]
kappa=Results[(c-1),3:4]
f=matrix(c(rep(0,n*k)),n,k)
for(j in 1:k){
f[,j]<-dvm(theta,mu[j],kappa[j])
}
tt<-matrix(c(rep(0,n*k)),n,k)
for(i in 1:n)
{
for(j in 1:k)
{
tt[i,j]<-(p[j]*f[i,j])/(p[1]*f[i,1]+p[2]*f[1,2])
}
}
for(j in 1:k)
} sorat<-sum(tt[,j]*sin(theta))
makhrj<-sum(tt[,j]*cos(theta))

```

```

    mu[j]<-atan(sorat/makhrj)
  if(makhrj<. )
  {
    mu[j]=pi+atan(sorat/makhrj)
  }else{
    mu[j]=ifelse(sorat>. ,atan(sorat/makhrj),۲*pi+atan(sorat/makhrj))
  }
}
Results[c,۱:۲]=c(mu[۱],mu[2])
d=c()
for(j in ۱:k)
{
  d=(sum(tt[,j]*cos(theta-mu[j])))/sum(tt[,j])
  kappa[j]<-((.۵۳-۱,۲۸*d^۲)*tan((pi*d)/2))
  Results[c,۳:۴]=c(kappa[۱],kappa[2])
  Results[c,۵]<-sum(tt[,۱])/n
  p=c(Results[c,۵],۱-Results[c,5])
  if (t(Results[c,]-Results[(c-۱),])%*%(Results[c,]-Results[(c-۱),])<. ,.....۱)(break)
  print(c)}
M[kk,]=Results[c,]
print(kk) }
M[kk,]
a<-c()
zz<-max.col(tt)
for(i in ۱:length(zz))
{
  if (zz[i]==۱) a=c(a,zz[i])
  i=i+۱
}
index=۱:n
index[zz==۱]
index[zz==2]
m=c(meanobs۱, meanobs۲, kappaobs۱, kappaobs۲, pobs)
MSE<-matrix(۱,nrow=۱,ncol=۵)
colnames(MSE)=c("MSE.mu۱", "MSE.mu۲", "MSE.kappa۱", "MSE.kappa۲",
"MSE.p1")
for (j in ۱:iteration)
{
  for (i in ۱:۵)
  {
    MSE[i]=(۱/j)*(sum((M[,i]-m[i])^۲))
  }
}
MSE

```

