مجله مدل‌سازی پیشرفته ریاضی، دوره ۴، شماره ۲، پاییز و زمستان ۱۳۹۳

برآورد شبه‌درستنمایی مدل کسینوسی توزیع فون‌میزس دومتغیره

سیما نوری جویباری و موسی گل‌علی‌زاده[۱]

گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۴/۱۱/۴ تاریخ پذیرش: ۱۳۹۴/۷/۱۸

چکیده: برای مدل‌بندی پدیده‌هایی که با زاویه شناسایی می‌شوند توزیع‌های جهتی ابزارهای بسیار مفیدی هستند. اخیرا، استفاده از این توزیع‌ها در علوم متنوعی مانند زیست‌شناسی، نجوم، هواشناسی و بیوانفورماتیک مورد توجه زیادی قرار گرفت. بویژه در تحقیقات علوم زیستی نشان داده شد که دو زوج زاویه وجود دارند که تا حد دقیقی ساختار هندسی و فضایی کامل یک پروتئین را در یک فضای سه بعدی توصیف می‌کنند. برای تشریح احتمالاتی براساس موقعیت اتم‌های پروتئین مقادیر توام این دو زاویه توزیعی به نام فون‌میزس دومتغیره وجود دارد. در این مقاله با مطالعه یکی از حالت‌های خاص این توزیع (مدل کسینوسی)، ابتدا به بررسی ویژگی‌های توزیع شامل تعداد مدهای توزیع و تقریب آن توسط نرمال دومتغیره پرداخته می‌شود. سپس نحوه برآورد پارامترهای توزیع به‌روش شبه‌درستنمایی تشریح می‌شود. مطالب نظری مقاله در مطالعه شبیه‌سازی بررسی و سپس کاربست مدل کسینوسی در یک مثال کاربردی ارزیابی می‌شود.

واژه‌های کلیدی: توزیع‌های دایره‌ای، توزیع فون‌میزس دو متغیره، برآورد شبه‌درستنمایی.

رده‌بندی ریاضی (۲۰۱۰): ۶۲H۱۱ و ۶۲P۱۰.

۱- مقدمه

در سالهای اخیر زاویه به عنوان داده آماری مورد توجه محققین بیشماری قرار گرفت. زاویه حرکت پرندگان، جهت وزش باد و جفت زوایای پیوندی بین آمینه اسیدهای یک پروتئین مثال‌هایی از این نوع داده هستند [۱]. یکی از جدیدترین کاربردهای آمار در تحقیقات علوم زیستی بررسی آماری توزیعی دومتغیره به نام فون‌میزس است که نقشی مشابه توزیع دومتغیره نرمال را بازی می‌کند. کاربرد این توزیع برای مدل‌بندی احتمالاتی ساختار هندسی پروتئین‌ها

۱- آدرس الکترونیکی نویسنده مسئول مقاله: golalizadeh@modares.ac.ir

توسط ماردیا و همکاران [۲] باعث شده اهمیت علمی آن بیش از حد شود. بویژه در سالیان متمادی، در نمایش هندسی دو زوج زاویه دوسطحی بین اتم‌های یک زنجیره پروتئین در نموداری به نام راماچاندران رسم می‌شود که اخیرا ماردیا و همکاران [۳] نشان دادند که نمایش معادل آن می‌تواند روی چنبره انجام گیرد. جالب اینجاست که توزیع فون‌میزس دومتغیره اولین بار [۴] توسط [۵] معرفی شد. اگرچه زوایای بی‌شماری را می‌توان برای پیوندهای گوناگون اتم‌ها درنظر گرفت، اما زوایای دوسطحی که رفتار تصادفی آنها توسط توزیع فون‌میزس دومتغیره توصیف می‌شود می‌تواند نقش ارزنده‌ای در پیشگویی ساختار پروتئین ایفا نمایند. واضح است که شناسایی ویژگی این توزیع و برآورد پارامترهای آن می تواند سنگ محکی برای استنباط‌های آماری مربوطه باشد. لذا، محققین علاقه‌مند به این موضوع سعی کردند حالت خاصی از توزیع فون‌میزس را درنظر بگیرند که تعداد پارامترهای کمتری دارد. به همین منظور سینگ و همکاران [۶] و ماردیا و همکاران [۷] به‌ترتیب مدل‌های سینوسی و کسینوسی را معرفی کردند. ماردیا و همکاران [۸] تعمیم چندمتغیره مدل سینوسی را نیز مطالعه کردند. اخیراً بررسی تک‌مدی بودن توزیع فون‌میزس چندمتغیره توسط ماردیا و وس [۹] انجام شد. بنابه منابع موجود بررسی ویژگی‌های متفاوت مدل کسینوسی کمتر مورد توجه قرار گرفت. مقاله حاضر سعی دارد به برخی از ویژگی‌های آماری این مدل بپردازد.

ساختار مقاله حاضر به‌صورت زیر تدوین شده است. در بخش دوم مدل کسینوسی توزیع فون میزس دومتغیره تعریف می‌شود. همچنین بررسی آن دسته از ویژگی‌های توزیع که در بدست آوردن برآورد پارامترها و نحوه تولید نمونه از آن مفید هستند، مدنظر قرار می‌گیرند. برآورد شبه‌درستنمایی پارامترهای توزیع در بخش سوم می‌آید. مطالعه شبیه‌سازی و مدل‌بندی کسینوسی زوایای دوسطحی در یک مثال کاربردی در بخش چهارم ارائه می‌شود. مقاله با بحث و نتیجه‌گیری تکمیل خواهد شد.

۲- خلاصه‌ای از مدل کسینوسی توزیع فون‌میزس دومتغیره

توزیع فون‌میزس یکی از توزیع‌های مفید و کارا برای مدل‌بندی داده‌های زاویه‌ای است. تحقیقات وسیعی راجع به حالت یک متغیره این توزیع صورت گرفته است که بخشی از آن در ماردیا و جاپ [۱] آمده است. حالت دومتغیره آن اخیرا مورد توجه محققین آمار کاربردی بویژه بیوانفورماتیک قرار گرفته است (به عنوان مثال [۱۰] را ببینید). از نقطه نظر تئوری، این توزیع ویژگی‌های جالبی دارد که در موضوعات کاربردی زیستی و بویژه بیوانفورماتیک مورد استفاده قرار خواهد گرفت. معرفی این توزیع و حالت‌های خاص آن در ادامه می‌آید.

از بین توزیع‌های موجود تک‌متغیره مدور برای مدل‌بندی زاویه، توزیع فون‌میزس تک‌متغیره نقش مهمی را ایفا می‌کند. این توزیع مشابه توزیع نرمال در فضای اقلیدسی است. اگر متغیر

زاویه‌ای θ دارای توزیع فون‌میزس تک‌متغیره با پارامترهای μ و κ باشد این گزاره را بصورت $\theta \sim VM(\mu, \kappa)$ نمایش می‌دهند و در این حالت تابع چگالی احتمال آن $f_\theta(\theta)$ بصورت

$$f(\theta)=\frac{1}{2\pi I_0(k)}e^{k\cos(\theta-\mu)}, \qquad 0\le\theta\le\pi$$

نوشته می‌شود که در آن پارامتر μ میانگین زاویه‌ای توزیع و κ پارامتر تمرکز است [۱]. عبارت $I_0(k)$ تابع بسل تعدیل یافته از مرتبه صفر است و بنابه آبراموویچ و استگان [۱۱] به صورت زیر تعریف می شود:

$$I_0(k)=\frac{1}{2\pi}\int_0^{2\pi}e^{\kappa cos\theta}d\theta=\sum_{r=0}^{\infty}(\frac{k}{2})^{2r}(\frac{1}{r!})^2 .$$

اکنون دو متغیر زاویه‌ای $-\pi\le\theta_1\le\pi$ و $-\pi\le\theta_2\le\pi$ را در نظر بگیرید. ماردیا [۵] یک تابع چگالی توام دو متغیره را برای جفت زاویه θ_1 و θ_2 به‌ازای $-\pi\le\mu_1,\mu_2\le\pi$ و $k_1,k_2\ge 0$ بصورت زیر تعریف کرد:

$$\begin{aligned} f(\theta_1,\theta_2)=(c(k_1,\,k_2,\,A))^{-1}\exp\{&k_1\cos(\theta_1-\mu_1)+\;k_2\cos(\theta_2-\mu_2)\\ &+[\cos(\theta_1-\mu_1),\sin(\theta_1-\mu_1)]\mathbf{A}[\cos(\theta_2-\mu_2),\sin(\theta_2-\mu_2)]^T .\end{aligned} \qquad (1)$$

در این تابع چگالی، k_1 و k_2 پارامترهای تمرکز و μ_1 و μ_2 به ترتیب میانگین زاویه‌ای متغیرهای زاویه‌ای θ_1 و θ_2 هستند. تابع $c(k_1, k_2, \mathbf{A})$ ثابت نرمال‌ساز و $\mathbf{A}$ ماتریسی معین مثبت با بعد ۲×۲ است. برای اینکه درک بهتری از تابع چگالی رابطه (۱) بدست آید می‌توان آن را برحسب ماتریس $\mathbf{A}$ بسط داد. فرض کنید به ازای $\mathbf{A}=[\mathbf{a_{ij}}]$ $i,j=1,2$ ماتریس موردنظر باشد. در آن صورت با جایگذاری اعضای این ماتریس در رابطه (۱) داریم:

$$\begin{aligned} f(\theta_1,\theta_2)=(c(k_1,k_2,A))^{-1}\exp\{&k_1\cos(\theta_1-\mu_1)+k_2\cos(\theta_2-\mu_2)\\ &+a_{11}\cos(\theta_1-\mu_1)\cos(\theta_2-\mu_2)+a_{21}\sin(\theta_2-\mu_2)\cos(\theta_2-\mu_2)\\ &+\;a_{12}\cos(\theta_1-\mu_1)\sin(\theta_2-\mu_2)+a_{22}\sin(\theta_1-\mu_1)\sin(\theta_2-\mu_2)\}. \end{aligned} \qquad (2)$$

ملاحظه می‌شود که تابع چگالی (۲) دارای ۸ پارامتر است، می‌توان انتظار داشت که وجود این تعداد پارامتر باعث پیچیده شدن محاسبات آماری بویژه بدست آوردن برآورد پارامترها می‌شود. مدل‌های مختلف با حضور پنج پارامتر به گونه‌ای که تقلیدی از توزیع نرمال دو متغیره باشد توسط[۶] و [۸] ارائه شد. اگر در ماتریس $\mathbf{A}$ قرار دهیم $a_{12}=a_{21}=0$ و $a_{11}=\alpha$ و $a_{22}=\beta$ آنگاه بدست می‌آوریم:

$$f(\theta_1,\theta_2)=(c(k_1,k_2,A))^{-1}\exp\{k_1\cos(\theta_1-\mu_1)+k_2\cos(\theta_2-\mu_2)$$
$$+\alpha\cos(\theta_1-\mu_1)\cos(\theta_2-\mu_2)+\beta\sin(\theta_1-\mu_1)\sin(\theta_2-\mu_2)\}.$$

حال با قرار دادن $\alpha=\beta=-k_3$ در این رابطه تابع چگالی مدل کسینوسی[1] توزیع فون‌میزس دومتغیره بصورت زیر حاصل خواهد شد:

$$f_{cosin}(\theta_1,\theta_2)=(c(k_1,k_2,k_3))^{-1}\exp\{k_1\cos(\theta_1-\mu_1)$$
$$+k_2\cos(\theta_2-\mu_2)-k_3\cos(\theta_1-\mu_1-\theta_2+\mu_2) \quad (۳)$$

که در آن

$$c(k_1,k_2,k_3) = C = (2\pi)^2\Big[I_0(k_1)I_0(k_2)I_0(k_3)+2\sum_{p=1}^{\infty}I_p(k_1)I_p(k_2)I_p(k_3)\Big] \quad (۴)$$

به قسمی که $k_1\geq k_2\geq k_3>0$ و $k_2>0$. واضح است که اگر در این مدل $k_3=0$ آنگاه متغیرهای زاویه‌ای θ_1 و θ_2 از هم مستقل خواهند بود لذا، می‌توان k_3 را پارامتر همبستگی نامید. علاوه بر این، اگر $k_1=k_2=k_3=0$ آنگاه توزیع حاشیه‌ای هرکدام از زوایا به توزیع یکنواخت روی دایره تبدیل می‌شود. بیان این نکته ضروری است که وقتی نوسانات θ_1 و θ_2 به اندازه کافی کوچک باشد (تغییرات دو زاویه نسبت به پارامترهای میانگین زاویه‌ایشان خیلی کم باشد)، آنگاه می‌توان توزیع مدل کسینوسی را با توزیع نرمال دو متغیره تقریب زد. برای اثبات این ادعا ابتدا می‌توان از تقریب $\cos(\theta_i-\mu_i)\approx 1-\frac{1}{2}\big[(\theta_i-\mu_i)^2\big]$ به‌ازای $i=1,2$ استفاده کرد. آنگاه با در نظر گرفتن $\mu_1=\mu_2=0$ محاسبات جبری مربوطه بصورت زیر است:

$$\begin{aligned}
f_{cosin}(\theta_1,\theta_2) &\propto \exp\big\{k_1\cos\theta_1+k_2\cos\theta_2-k_3\cos(\theta_1-\theta_2)\big\}\\
&\approx \exp\Big\{k_1\big(1-\frac{\theta_1^2}{2}\big)+k_2\big(1-\frac{\theta_2^2}{2}\big)-k_3\big(1-\frac{(\theta_1-\theta_2)^2}{2}\big)\Big\}\\
&= \exp(k_1+k_2-k_3)\exp\Big\{\frac{-1}{2}[k_1,\theta_1^2+k_2,\theta_2^2-k_3(\theta_1-\theta_2)^2]\Big\}\\
&\propto s_2^*\exp\Big\{\frac{-1}{2}[\theta_1^2(k_1-k_3)+\theta_2^2(k_2-k_3)+2k_3\theta_1\theta_2]\Big\}\\
&= s_2^*\exp\{\frac{-1}{2}(\theta_1,\theta_2)\begin{pmatrix}k_1-k_3 & k_3\\ k_3 & k_2-k_3\end{pmatrix}(\theta_1,\theta_2)^T\}\\
&= s_2^*\exp\Big\{\frac{-1}{2}\Theta^T\Sigma^{-1}\Theta\Big\}
\end{aligned} \quad (۵)$$

1- Cosine model

که $\Theta=(\theta_1,\theta_2)^T$ و $\Sigma^{-1}=\begin{pmatrix} k_1-k_3 & k_3 \\ k_3 & k_2-k_3 \end{pmatrix}$ و s_2^* ثابت نرمال‌ساز است. واضح است اگر Σ^{-1} معین مثبت باشد تقریب حاصل معتبر خواهد بود. از این رو، علاوه بر این که باید عناصر روی قطر مثبت باشند یعنی $(k_2-k_3)>0$ و $(k_1-k_3)>0$ باید $\left|\Sigma^{-1}\right|>0$ یعنی $(k_1-k_3)(k_2-k_3)>k_3^2$ یا به‌طور معادل $k_3<\frac{k_1k_2}{k_1+k_2}$. پس به‌طور خلاصه، اگر نوسانات دو زاویه θ_1 و θ_2 خیلی کم باشد آنگاه مدل کسینوسی (۳) را می‌توان توسط توزیع نرمال دومتغیره با بردار میانگین $(\mu_1,\mu_2)^T$ و ماتریس کواریانس $\Sigma=\frac{1}{(k_1-k_3)(k_2-k_3)}\begin{pmatrix} k_1-k_3 & k_3 \\ k_3 & k_2-k_3 \end{pmatrix}$ تقریب زد (قضیه ۱، [۷]).

بعد از اینکه خواص چگالی توام برای بردار $(\theta_1,\theta_2)^T$ مطالعه شد، بررسی توابع چگالی حاشیه‌ای نیز می‌تواند جالب توجه باشد. برای بدست آوردن تابع چگالی حاشیه‌ای θ_2 با فرض اینکه $\mu_1=\mu_2=0$ ابتدا با پیروی از ماردیا و همکاران [۷]، فرض کنید توابع $\frac{-\pi}{2}\le\beta(\theta_2)=\beta\le\frac{\pi}{2}$ و $\alpha(\theta_2)=\alpha$ به‌صورت زیر تعریف شوند:

$$\tan(\beta(\theta_2))=\frac{-k_3\sin\theta_2}{k_1+k_3\cos\theta_2} \quad \text{و} \quad \alpha(\theta_2)=\sqrt{k_1^2+k_3^2-2k_1k_3\cos\theta_2} \qquad (۶)$$

حال، با توجه به تعریف تابع چگال مدل کسینوسی (رابطه (۳) را ببینید) و تساوی‌های (۶) چگالی حاشیه‌ای θ_2 در مدل کسینوسی با استفاده از انتگرال‌گیری تابع چگالی توام $(\theta_1,\theta_2)^T$ نسبت به متغیر زاویه‌ای θ_1 به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned} f_{cosin}(\theta_2)&=\int_{-\pi}^{\pi} f_{cosin}(\theta_1,\theta_2)d\theta_1 \\ &=\int_{-\pi}^{\pi}\{C\}^{-1}\exp\left\{k_1\cos\theta_1+k_2\cos\theta_2-k_3\cos(\theta_1-\theta_2)\right\}d\theta_1 \qquad (۷)\\ &=\{C\}^{-1}2\pi I_0(\alpha(\theta_2))\exp\{k_2\cos\theta_2\} \end{aligned}$$

که در آن $I_0(q)$ تابع بسل از مرتبه صفر است. اکنون، با داشتن توزیع حاشیه‌ای θ_2 می‌توان توزیع شرطی θ_1 به شرط θ_2 را بدست آورد. با محاسبات جبری ساده داریم:

$$\begin{aligned} f_{cosin}(\theta_1|\ \theta_2)&=\frac{f_{cosin}(\theta_1,\theta_2)}{f_{cosin}(\theta_2)} \\ &=(2\pi I_0(\alpha(\theta_2)))^{-1}\exp\{\alpha(\theta_2)\cos(\theta_1-\beta(\theta_2))\}. \end{aligned}$$

لذا، تابع چگالی شرطی $f_{\cos in}(\theta_1,\theta_2)$ توزیع فون‌میزس تک‌متغیره با پارامتر تمرکز $\alpha(\theta_2)$ و میانگین دایره‌ای $\beta(\theta_2)$ است.

علاوه بر خواصی که تاکنون بدست آمد ویژگی‌های دیگری نیز برای مدل کسینوسی وجود دارد که بررسی آنها ارزشمند خواهد بود. تعدادی از آنها در قالب قضیه در ذیل آمده است. جزئیات بیشتر این قضایا را می‌توان در ماردیا و همکاران [۷] مطالعه کرد. توجه شود که قضایای ذیل کمک شایانی به محاسبه برآورد پارامترها به روش گشتاوری و ماکسیمم درستنمایی و همچنین نحوه تولید نمونه از مدل کسینوسی می‌کند.

قضیه ۱: در تابع چگالی مدل کسینوسی و به ازای $k_2 > k_3 > 0$ و $k_1 > k_3 > 0$ اگر $k_3 < \frac{k_1 k_2}{k_1 + k_2}$ آنگاه توزیع مربوطه تک‌مدی و در غیر اینصورت دومدی است.

قضیه ۲: مدل کسینوسی با $k_3 \neq 0$ و $\mu \neq 0$ را در نظر بگیرید. آنگاه تابع چگالی حاشیه‌ای θ_2 حول μ_2 متقارن و تک‌مدی با مد μ_2 است اگر و فقط اگر $A(|k_1 - k_3|) \leq \frac{|k_1 - k_3|\, k_2}{k_1 k_3}$. در غیر اینصورت تابع چگالی حاشیه‌ای دومدی با مدهای $\mu_2 - \theta_2^*$ و $\mu_2 + \theta_2^*$ است که در آن θ_2^* از حل تساوی $b(\theta_2^*) = 0$ بدست می‌آید، طوریکه $b(\theta_2^*) = -k_2 + \frac{k_1 k_3 A(\alpha(\theta_1^*))}{\alpha(\theta_1^*)}$ و $\alpha(\theta_2) = \sqrt{k_1^2 + k_3^2 - 2k_1 k_3 \cos\theta_2}$.

۳- برآورد شبه‌درستنمایی پارامترهای توزیع

روش برآورد شبه‌درستنمایی[1] (PL) یکی از روش‌های بسیار قوی برای برآورد پارامترهاست که مفهوم آن اولین‌بار توسط بیساگ [۱۲] معرفی شده است. برآورد حاصل از این روش به دلیل محاسبات کمترش می‌تواند به‌عنوان جایگزین برآورد ماکسیمم درستنمایی[2] (ML) درنظر گرفته شود. اخیرا ماردیا و همکاران [۱۳] روش برآورد PL را برای خانواده‌های نمایی مورد مطالعه قرار دادند. در واقع آنها نشان دادند برآورد PL پارامترهای خانواده‌های نمایی که بصورت حاشیه‌ای بسته‌اند را می‌توان با برآوردهای بدست آمده از روش ML برابر دانست.

بر اساس نمونه تصادفی به حجم n از متغیر تصادفی از بردار P بعدی $\Theta = (\theta_1, \theta_2, ..., \theta_p)^T$ تابع شبه‌درستنمایی عبارتست از:

1- Pseudo Likelihood
2- Maximum Likelihood

$$\mathrm{PL}=\prod_{j=1}^{p}\prod_{i=1}^{n} f_j(\theta_{ji}|\ (\theta_{1i},\theta_{2i},...,\theta_{pi});q)$$

به قسمی‌که $f_j(\theta_{ji}|\ ...;q)$ توزیع شرطی بوده و به پارامتر θ_j بستگی دارد و θ_{ji}، i–امین مشاهده از j–امین متغیر و q پارامتر مجهول با طول r است.

از آنجایی‌که مطالعه مورد نظر معطوف به زوج زاویه θ_1 و θ_2 است تابع درستنمایی مربوطه براساس بردار مشاهدات $(\theta_{1i},\ \theta_{2i})^T$ به‌ازای $i=1,...,n$ بصورت

$$\mathrm{PL}=\prod_{i=1}^{n} f_{cosin}(\theta_{1i}|\ \theta_{2i},\ q)f_{cosin}(\theta_{2i}|\ \theta_{1i},\ q)$$

خواهد بود که در آن $f_{cosin}(\theta_{1i}|\ .,q_{cosin})$ تابع چگالی شرطی مورد نیاز و q_{cosin} پارامترهای آن است. بنابه [۱۳] مدل‌های سینوسی و کسینوسی فون‌میزس دومتغیره، مدل‌های نمایی نابسته‌اند. اما می‌توان آنها را با مدل‌های نمایی تقریب زد تا برآوردهای حاصل از ماکسیمم کردن لگاریتم تابع شبه‌درستنمایی تقریبی از برآوردهای حاصل از روش ML شوند. جزئیاتی از روش محاسبات برای مدل کسینوسی در ادامه تشریح می‌شود.

از بخش قبل می‌دانیم که توزیع شرطی در مدل کسینوسی توزیع فون‌میزس تک‌متغیره است. بنا به نمادگذاری این بخش می‌نویسیم: $q_{cosin}=(\mu_1,\ \mu_2,\ k_1,\ k_2,k_3)^T$.

با فرض اینکه $\mu_1=\mu_2=0$ تابع شبه‌درستنمایی مدل کسینوسی برابر

$$\begin{aligned} PL_{cosin} &= \prod_{i=1}^{n} f(\theta_{1i}|\ \theta_{2i},q)f(\theta_{2i}|\ \theta_{1i},q) \\ &= \prod_{i=1}^{n}\left\{\frac{\exp\left[\alpha(\theta_{2i})\cos(\theta_{1i}-\beta(\theta_{2i}))\right]}{2\pi I_0(\alpha(\theta_{2i}))}\right\}\cdot\left\{\frac{\exp\left[\alpha(\theta_{1i})\cos(\theta_{2i}-\beta(\theta_{1i}))\right]}{2\pi I_0(\alpha(\theta_{1i}))}\right\} \end{aligned} \quad (8)$$

است که در آن به ازای $i=1,...,n$،

$$\tan\beta(\theta_{2i})=\frac{-k_3}{k_1-k_3\cos\theta_{2i}},\quad \alpha(\theta_{2i})^2=k_1^2+k_3^2-2k_1k_3\cos\theta_{2i}$$

$$\tan\beta(\theta_{1i})=\frac{-k_3}{k_2-k_3\cos\theta_{1i}},\quad \alpha(\theta_{1i})^2=k_2^2+k_3^2-2k_2k_3\cos\theta_{1i}.$$

بنابه هملریک و همکاران [۱۰] می‌توان شکل ساده‌تر PL_{cosin} را به‌صورت

$$PL_{cosin}=\left\{\prod_{i=1}^{n}\frac{\exp\left[k_1\cos\theta_{1i}-k_2\cos(\theta_{1i}-\theta_{2i})\right]}{2\pi I_0(\alpha(\theta_{1i}))}\right\}\times\left\{\frac{\exp\left[k_2\cos\theta_{2i}-k_2\cos(\theta_{1i}-\theta_{2i})\right]}{2\pi I_0(\alpha(\theta_{2i}))}\right\}$$

نوشت. برای محاسبه برآورد شبه‌درستنمایی پارامترها باید تابع شبه‌درستنمایی (۸) یا لگاریتم آن ماکسیمم شود. مشتق‌گیری از لگاریتم تابع درستنمایی (PL_{cosin}) نسبت به پارامترهای مدل و مساوی صفر قرار دادن آنها منجر به تساوی‌های

$$\frac{\partial\log(PL_{cosin})}{\partial k_1}=\sum_{i=1}^{n}\cos\theta_{1i}-\sum_{i=1}^{n}\left[\frac{k_1-k_2\cos\theta_{2i}}{\alpha(\theta_{2i})}\right]A_1(\alpha(\theta_{2i}))=0$$

$$\frac{\partial\log(PL_{cosin})}{\partial k_2}=\sum_{i=1}^{n}\cos\theta_{2i}-\sum_{i=1}^{n}\left[\frac{k_2-k_2\cos\theta_{1i}}{\alpha(\theta_{1i})}\right]A_1(\alpha(\theta_{1i}))=0$$

$$\frac{\partial\log(PL_{cosin})}{\partial k_2}=-2\sum_{i=1}^{n}\cos(\theta_{1i}-\theta_{2i})-\sum_{i=1}^{n}\left[\frac{k_2-k_1\cos\theta_{2i}}{\alpha(\theta_{2i})}\right]A_1(\alpha(\theta_{2i}))$$
$$-\sum_{i=1}^{n}\left[\frac{k_2-k_2\cos\theta_{1i}}{\alpha(\theta_{1i})}\right]A_1(\alpha(\theta_1))=0$$

می‌شود که در آن $A_p(x)=\frac{I_p(x)}{I_0(x)}$. می‌توان پیش‌بینی کرد که حل این دستگاه معادلات برحسب پارامترهای k_1، k_2 و k_2 منجر به جواب‌های صریحی برای پارامترهای مورد اشاره نشود. لذا، استفاده از روش‌های محاسباتی بهینه‌سازی اجتناب‌ناپذیر است. این رویکرد در مطالعه شبیه‌سازی و همچنین مثال کاربردی به‌منظور ارزیابی برآورد PL پارامترهای مدل کسینوسی مدنظر قرار می‌گیرد. قابل اشاره است که برای محاسبه برآور د PL از دستور optim (همراه با اعمال برخی محدودیت‌های مربوط به مسئله مورد بررسی در این مقاله) در زبان برنامه‌نویسی R استفاده نمودیم.

۴- نمونه‌گیری از مدل کسینوسی

در این بخش ابتدا برای ارزیابی برخی از ویژگی‌های معرفی شده یک مطالعه شبیه‌سازی انجام خواهد شد. سپس، با درنظر گرفتن مجموعه‌ای از داده واقعی بعضی از مطالب تشریح شده در بخش‌های قبل در قالب مثال کاربردی مورد کنکاش قرار می‌گیرد. روش‌های متنوعی برای شبیه‌سازی آماری این زوایا وجود دارد [۱۴]. در این بخش به نحوه تولید نمونه تصادفی زوایای دوسطحی براساس مدل کسینوسی پرداخته می‌شود.

بنابه مطالب ارائه شده در بخش قبل، توزیع شرطی θ_1 به شرط θ_2 دارای توزیع فون‌میزس

تک‌متغیره است. لذا بنا به ماردیا و همکاران [۷] می‌توان شبیه‌سازی از این توزیع شرطی را با شبیه‌سازی از توزیع حاشیه‌ای درهم آمیخت و زوج نمونه‌های تصادفی برای بردار $(\theta_1,\theta_2)^T$ بدست آورد. برای یکنواخت‌سازی نمادها فرض کنید تابع چگالی حاشیه‌ای $f(\theta_2)$ و تابع چگالی شرطی $f(\theta_1|\ \theta_2=\theta_2)$ موجود باشند.

باید توجه داشت که شبیه‌سازی مستقیم از تابع چگالی حاشیه‌ای $f(\theta_2)$ میسر نیست. با این حال، با وجود روش‌های بی‌شمار شبیه‌سازی از توزیع‌های آماری می‌توان از نمونه‌گیری رد و پذیرش[1] برای تولید نمونه از این چگالی حاشیه‌ای استفاده کرد [۱۵]. در الگوریتم رد و پذیرش، باید توزیعی نامزد به‌گونه‌ای انتخاب شود که هم‌رفتار و دارای توزیعی با دامنه تعریف یکسان با توزیع موردنظر بوده و تابع چگالی هدف در زیر تابع چگالی کاندید قرار گیرد [۱۵]. برای رسیدن به این مهم در بخش‌های قبل نحوه شناسایی تک یا دومدی بودن تابع چگالی حاشیه‌ای با استفاده از برخی قضایا ارائه شد. علاوه بر آن‌ها، ماردیا و همکاران [۷] توجه به دو نکته ذیل را توصیه کردند:

- اگر تابع چگالی حاشیه‌ای θ_2 تک‌مدی است توزیع $VM(\mu_1,\kappa^*)$ به عنوان توزیع نامزد انتخاب شود.
- در صورتی که تابع چگالی حاشیه‌ای دومدی است از آمیختگی با وزن‌های یکسان دو توزیع فون‌میزس $VM(\mu_1+\theta_2^*,\kappa^*)$ و $VM(\mu_1-\theta_2^*,\kappa^*)$ به عنوان تابع نامزد انتخاب شود طوریکه مقدار θ_2^* ریشه تابع $b\left(\theta_2^*\right)=0$ مورد اشاره در قضیه ۲ است. بعلاوه، باید κ^* به گونه‌ای انتخاب شود که اختلاف ناچیزی بین تابع چگالی نامزد و تابع چگالی حاشیه‌ای $f(\theta_2)$ وجود داشته باشد.

با توجه به این دونکته، اجرای الگوریتم رد و پذیرش به آسانی امکان‌پذیر است. برای نیل به این هدف پیروی از چند گام مرسوم شبیه‌سازی ضروری است که عبارتند از:

گام اول) از توزیع نامزد با تابع چگالی $h(\theta_2)$ داده تولید شود.

گام دوم) به ازای هر $-\pi\le\theta_2\le\pi$ ثابت L بصورت زیر در نظر گرفته شود:

$$L=max\left\{\frac{f(\theta_2)}{h(\theta_2)},1\right\}.$$

گام سوم) یک نمونه از $U(0,1)$ تولید شود.

1- Accepted Rejected Sampling

گام چهارم) اگر نامساوی $U < \frac{f(\theta_2)}{L \times h(\theta_2)}$ برقرار بود نمونه تولیدی در **گام اول** به‌عنوان نمونه‌ای تصادفی از توزیع $f(\theta_2)$ پذیرفته شود و درغیر اینصورت به **گام اول** رجوع شود.

اکنون با داشتن نمونه‌ای برای θ_2، به‌منظور تولید نمونه تصادفی θ_1 می‌توان از توزیع شرطی $f(\theta_1 | \theta_2 = \theta_2)$ که دارای توزیع فون‌میزس با پارامترهای تمرکز $\alpha(\theta_2)$ و میانگین $\beta(\theta_2)$ است کمک گرفت. توجه کنید که زوج $\{\alpha(\theta_2), \beta(\theta_2)\}$ پارامترهایی هستند که در بخش ۳ معرفی شدند.

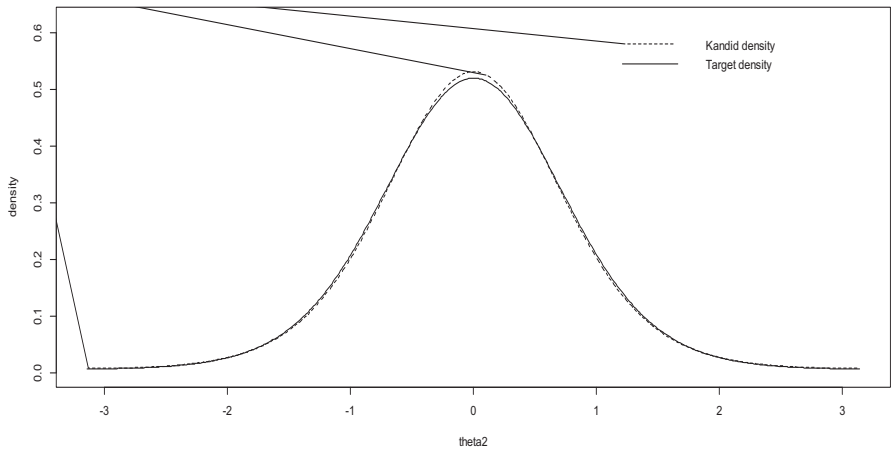
۵- شبیه‌سازی و تحلیل مثال‌کاربردی

به‌کمک الگوریتم رد و پذیرش ۱۰۰۰ نمونه از مدل کسینوسی توزیع فون‌میزس دومتغیره وقتی که $\mu_1 = \mu_2 = 0$ اما با پارامترهای متفاوت دیگر تولید و فرآیند مورد اشاره برای بررسی دقیق از روند شبیه‌سازی، ارزیابی و نحوه عملکرد روش‌های برآورد فرآیند شبیه‌سازی ۱۰۰ بار تکرار شد. بیان این نکته ضروری است که درصد پذیرش نمونه بطور متوسط ۶۰ بود. شکل ۱ نشان می‌دهد که تابع چگالی هدف در زیر تابع چگالی کاندید یعنی فون‌میزس تک متغیره قرار گرفته است. به‌طور خلاصه، نمونه‌های تولید شده در این مطالعه شبیه‌سازی در ماتریسی با بعد ۲۰۰× ۱۰۰۰ ذخیره و روش‌های متفاوت برآورد به‌کار گرفته شد. شکل خلاصه‌شده‌ای از نتایج در جدول ۱ آمده است. توجه کنید که ابتدا نمونه‌ها از مدل کسینوسی با پارامترهای $(\mu_1, \mu_2, k_1, k_2, k_3) = (0, 0, 4, 3, 1)$ شبیه‌سازی شد. سپس با مجهول دانستن پارامترها برآورد آنها محاسبه و اریبی، خطای برآورد و فواصل اطمینان تقریبی نرمال q درصدی محاسبه شدند. فاصله اطمینان مورد نظر بصورت $T \pm \frac{\sigma_I}{\sqrt{n}} Z_q$ بدست آمده که در آن آماره T میانگین n بار (تعداد تکرار) برآورد پارامتر مورد نظر است.

ابتدا، توجه‌مان را معطوف به اولین پارامتر مدل (k_1) می‌کنیم. قبل از شبیه‌سازی مقدار آن در عدد ۴ ثابت درنظر گرفته شد. همانگونه که از جدول ۱ ردیف مربوط به این پارامتر ملاحظه می‌شود مقدار برآورد PL برابر ۴/۰۷۵ (با خطای برآورد۰/۵۹۶) و فاصله اطمینان تقریبی (۳/۹۵۸,۴/۱۹۲) بدست آمده است. به‌نظر می‌رسد این روش توانسته برآورد قابل قبولی برای پارامتر ارائه k_1 کند.

قبل از اجرای شبیه‌سازی، برای مقدار پارامتر تمرکز دوم (k_2) عدد ثابت۳ درنظر گرفته شد. برآورد حاصل از روش PL برای این پارامتر برابر ۳/۰۶۹ (با خطای برآورد ۰/۵۲۹) با فاصله اطمینان تقریبی (۲/۹۶۵, ۳/۱۷۳) بدست آمده است. با کمک نتایج حاصل، می‌توان نتیجه

گرفت که این روش برآورد به مقدار واقعی بسیار نزدیک است.



شکل۱: نمودار تابع کاندید (نقطه‌چین) و تابع هدف (ممتد)

جدول ۱: نتایج حاصل از برآورد به روش PL پارامترهای مدل کسینوسی $\mu_1 = \mu_2 = 0$ با مقادیر واقعی پارامترها به‌صورت $k_1 = 4$، $k_2 = 3$، $k_3 = 1$

پارامتر	برآورد	اریبی	خطای برآورد	فاصله اطمینان	
				حد پایین	حد بالا
k_1	۴/۰۷۵	۰/۰۷۵	۰/۵۹۶	۳/۹۵۸	۴/۱۹۲
k_2	۳/۰۶۹	۰/۰۶۹	۰/۵۲۹	۲/۹۶۵	۳/۱۷۳
k_3	۰/۹۹۹	۰/۰۰۱	۰/۴۲۰	۰/۹۱۶	۱/۰۸۱

توجه به پارامتر k_3 اهمیت قابل توجهی در میزان همبستگی دارد. مقدار این پارامتر قبل از اجرای شبیه‌سازی مقدار ثابت۱ در نظر گرفته شده است. برآوردهای حاصل از روش PL برای این پارامتر برابر ۱/۱۹۹ با خطای برآورد۴۲۰ /۰ و با فاصله اطمینان تقریبی ۹۵ درصدی (۰/۹۱۶,۱/۰۸۱) بدست آمده است که نشان می‌دهد برآورد حاصل به مقدار واقعی پارامتر بسیار نزدیک است.

زوایای دو سطحی[1] به عنوان یکی از مهمترین مختصات درونی در تعیین ساختار هندسی

1- Dihedral angels

پروتئین بحساب می‌آیند، که بصورت زوج زوایای (ϕ,ψ) نمایش داده می‌شود [۲]. ساختار دوم پروتئین از دوران زوایای (ϕ,ψ) بوجود می‌آید. به طور دقیق‌تر، زاویه بین نیتروژن و کربن آلفا زاویه‌ی ϕ و زاویه‌ی بین پیوند کربن آلفا و کربن کربونیل زاویهی ψ درنظر گرفته می‌شود [۱۶]. اما زاویه‌ی دیگری که حول پیوند پپتیدی وجود دارد زاویه‌ی ω نامیده می‌شود که جز با ثابت نگه داشته شدن در دو مقدار خاص نقش دیگری در تعیین ساختار هندسی پروتئین ایفا نمی‌کند. توجه شود که اندازه این دو زاویه در هر بخش از پروتئین در یک بازه‌ی زمانی خاص در طول چرخش متفاوت هستند. مطالعه آماری این زوایا با رویکرد مدل‌بندی به موضوع پیش بینی ساختار پروتئین‌ها سابقه طولانی دارد. از سالیان دور دانشمندان علوم زیستی پراکندگی نقاط مشاهده شده بر روی چنبره[1] را در یک صفحه رسم می‌کنند که به نمودار حاصل نمودار راماچاندران[2] گفته می‌شود [۱۷]. در واقع نمودار راماچاندران وضعیت مجاز هر زاویه را برای ساختارهای پروتئین نشان می‌دهد. ماردیا و همکاران [۱۸] با استفاده از این ایده که توزیع یک زاویه بر روی دایره و توزیع توام دو زاویه دربازه ($-\pi,\pi$) بر روی چنبره قرار می‌گیرد به مدل بندی آماری این زوایا بر روی چنبره پرداختند. چرخش یک بخش کوچک از پروتئین که دربرگیرنده این دو زوج زاویه است مجموعه‌ای از زوایای دوسطحی بصورت $(\phi_1,\psi_1),(\phi_2,\psi_2),...,(\phi_n,\psi_n)$ را در اختیار محقق قرار می‌دهد. اگرچه از نقطه نظر زیستی انواع و اقسام همبستگی بین هر دو زوج نمونه قابل تصور است، اما فرض ما بر این است که نمونه به حجم n از این زوایا در واقع نمونه‌ای تصادفی از یک توزیع فون‌میزس دومتغیره است.

اسیدهای آمینه را براساس خصوصیتی که دارند به گروه‌های مختلف تقسیم می‌کنند. آلانین ALA از جمله اسیدهای آمینه‌هایی است که در تمام پروتئینها موجود و از لحاظ تقسیم بندی در گروهی از اسیدآمینه‌های غیر ضروری که دارای زنجیر جانبی خطی هستند قرار می گیرد. همچنین آلانین در گروهی از اسیدهای آمینه R ناقطبی قرار می‌گیرد. در این مقاله داده‌هایی که مورد مطالعه قرار می‌گیرند با استفاده از نرم‌افزار YASARA از پروتئین آلفا۱-آنتی تریپسین (AAT) که متشکل از ۳۹۳ اسید آمینه است و هر اسید آمینه دارای زوایای دوسطحی ϕ و ψ می‌باشد، بدست آمده است. در ساختار این پروتئین صفحات بتا و پیچ‌های آلفا وجود دارند که در آزمایشگاهی زیستی به تعداد ۱۰۰۰زاویه ϕ و ψ از اسیدآمینه ALA۵۸ و ALA۲۴۸ شبیه‌سازی شد. اسید آمینه ALA۵۸ بر روی پیچ‌های آلفا قرار دارد که دارای تحرک کمتری نسبت به اسیدآمینه ALA۲۴۸، که بر روی صفحات بتا قرار دارد، است. بدین جهت زوایای موجود در اسید آمینه ALA۵۸ را سخت $(Rigid)$ و زوج زوایای موجود در اسیدآمینه۲۴۸ ALA را انعطاف‌پذیر $(Flex)$ نامیم. در نرم‌افزار آماری R با

1- Torus
2- Ramachandaran

استفاده از دستور *circ.cor* در کتابخانه ی *CircStat* مقدار آماره‌ی آزمون، $P-$ مقدار و مقدار همبستگی قابل محاسبه است. آزمون همبستگی را هم برای داده‌های *Rigid* و هم *Flex* بکار بردیم. مقدار آماره‌ی آزمون و نتایج حاصل از آن در جدول۲ آمده است. نتایج حاصل از این جدول نشان می‌دهد که زوایای دوسطحی برای هردو مجموعه داده همبستگی معنی‌داری ندارند و در این صورت می‌توان نمونه‌ها را مستقل از هم درنظر گرفت.

جدول ۲: نتایج آزمون همبستگی برای داده‌های واقعی

نوع اسید آمینه	زاویه	P–مقدار	مقدار همبستگی	مقدار آماره آزمون
ALA۵۸	ϕ	۰/۸۷۵	۰/۰۰۱	۰/۱۵۶
ALA۵۸	ψ	۰/۳۸۰	-۰/۰۰۸	-۰/۸۷۷
ALA۲۴۸	ϕ	۰/۳۱۳	-۰/۰۱۰	-۱/۰۰۷
ALA۲۴۸	ψ	۰/۸۵۷	۰/۰۰۱	۰/۱۷۹

اکنون می‌توان مطالب مربوطه برآورد پارامترها را برای این داده‌ها بکار گرفت. با پیروی از ارائه شده نتایج حاصل از برآورد پارامترهای دادهای انعطاف‌پذیر در جدول۳ آمده است. ملاحظه می‌شود که مقادیر برآورد برای پارامترهای k_1 و k_2 با روش برآورد PLبه‌ترتیب ۰/۰۵۳، ۰/۰۶۱ و ۰/۰۰۴ برای پارامتر k_3 بدست آمده است. به‌طور مشابه برای داده‌های سخت برآوردهای این پارامترها با روش برآورد PL به ترتیب برابر ۰/۰۰۴، ۰/۰۱۹ و ۰/۰۰۰ شد.

جدول ۳: نتایج حاصل از برآورد پارامترهای مدل کسینوسی از توزیع فون‌میزس دومتغیره برای داده‌های واقعی

	Rigid	*Flex*
k_1	۰/۰۰۴	۰/۰۵۳
k_2	۰/۰۱۹	۰/۰۶۱
k_3	۰/۰۰۰	۰/۰۰۴

۶- بحث و نتیجه‌گیری

مطالعه برخی از پدیده‌های زیستی در سالیان اخیر امری ضروری شده است. از بین آن‌ها شناسایی ساختارهای متفاوت پروتئین به‌دلیل کمک به درمان بعضی از بیماری‌ها نقش مهمی تری دارد. علوم آماری توانسته‌اند مدل‌بندی احتمالاتی پدیده‌های تصادفی مربوط به تغییرات

اتم‌های پروتئین نقش بسزایی در این امر مهم ایفا کند. توزیع فون‌میزس دومتغیره به دلیل فرم تابعی خاصش می‌تواند برای توصیف تغییرات زاویه دوسطحی بسیار مناسب باشد. لذا، رویکرد آماری به مطالعه ویژگی این توزیع از هر دو جنبه نظری و کاربردی مفید فایده خواهد بود. هدف مقاله حاضر مطالعه یکی از حالت‌های خاص توزیع فون‌میزس دومتغیره بود. برای این منظور ویژگی توزیع، نحوه برآورد پارامترهای آن و شیوه اخذ نمونه تصادفی ارائه شد. در مطالعه شبیه‌سازی، عملکرد برآورد پارامترها به روش PL مورد ارزیابی قرار گرفت. نحوه کاربست روش‌ها درمورد داده‌های واقعی نیز ارائه شد. جنبه‌های استنباطی دیگر مانند محاسبه نواحی اطمینان دقیق برای پارامترها، پیاده‌سازی روش‌های بیزی در مدل کسینوسی می‌تواند مسیرهایی برای تحقیقات آتی باشد.

مراجع

[1] Mardia, K. V and Jupp, P. (2000), *Directional Statistics*, John Wiley and Sons, New York.

[2] Mardia, K. V., Taylor, C. C. and Subramanian, G. K. (2003), Applications of Circular Distribution to Conformational Angles in Proteins, In *Proceedings of the 22nd LASR Workshop,* edited by KV Mardia, RG Aykroyd and MJ Langdon*,* 149-152. Leeds University Press.

[3] Mardia, K. V., Kent, J. T. and Taylor, C. C. (2010), Matching Unlabelled Configurations and Protein Bioinformatics. *Research Report STAT10-01, University of Leeds.*

[4] Mardia, K. V. (1972), *Statistics of Directional Data*, Academic Press, London.

[5] Mardia, K. V. (1975), Statistics of Directional Data, *Journal of the Royal Statistical Society Series B (Methodological),* **37**, 349-393.

[6] Singh, H., Hnizdo, V. and Demchuk, E. (2002), Probabilistic Model for Two Dependent Circular Variable, *Bioinformatics*, **89**, 719-723.

[7] Mardia, K. V., Hughes, G., Taylor, C. C. and Subramanian. G. K. (2007b), Bivariate Von Mises Densities for Angular Data with Application to Protein Bioinformatics, *Annals of Statistics*, **35**, 166-180.

[8] Mardia, K.V., Hughes, G., Taylor, C.C. and Singh, H. (2008). A Multivariate Von Mises Distribution with Applications to Bioinformatics, *Canadian Journal of Statistics*, **36**, 99-109.

[9] Mardia, K. V. and Voss, J. (2014), Some Fundamental Properties of a Multivariate Von Mises Distribution, *Communications in Statistics-Theory and Methods*, **43**, 1132-1144.

[10] Hamelryck, T., Mardia, K. V. and Ferkinghoff-Borg, J. (2012), *Bayesian Methods in Structural Bioinformatics, Statistics for Biology and Health*, Springer-Verlag, Heidelberg.

[11] Abramowitz, M and Stegun, I. A. (1965), *Handbook of Mathematical Functions*, Dover.

[12] Besag, J. (1975), Statistical Analysis of Non-Lattice Data, *The Statistician*, **24**, 179-195.

[13] Mardia, K. V., Kent, J. T., Hughes, G. and Taylor, C. C. (2009), Maximum Likelihood Estimation Using Composite Likelihoods for Closed Exponential Families, *Biometrika*, 21-30.

[14] Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer, New York.

[15] Ross, S, M. (2006), *Simulation*, (4th edition), Academic Press, Amsterdam.

[16] Bourne E. B. and Weissig H. E. (2009), *Structural Bioinformatics*, (2nd edition), John Wiley and Sons, New Jersey.

[17] Ramakrishan, C. and Ramachandran, G. N. (1965). Stereochemical Criteria for Polypeptide and Protein Chain Conformation ,*Biophysical Journal,* ,۵ 909-933.

[18] Mardia, K. V., Taylor, C. C. and Subramanian, G. K. (2007a), Protein Bioinformatics and Mixtures of Bivariate Von Mises Distribution for Angular Data, *Biometrics*, **63**, 502-512.

Pseudo-likelihood Estimator of the Bivariate Von-Mises Cosine Model

Sima Nouri Jouybari, Mousa Golalizadeh

Department of Statistics, Tarbiat Modares University, Tehran, Iran

Abstract

Directional statistics are very useful tools to model the phenomenon that are characterized by the angles. Recently, various disciplines including biology, astronomy, meteorology and bioinformatics have paid attention to use these distributions. Particularly, it was shown in biological researches that there are two pair angles describing, relatively, the complete geometrical and spatial structures of a protein in the three dimensional space. There is a distribution, called bivariate Von-Mises, to represent the position of the atoms based upon the values of these angles in a probabilistic manner. In this paper, considering an especial case of this density (cosine model), the properties of distribution including the numbers of modes and its approximation by the bivariate normal distribution are first studied. Then, to estimate the parameters using the pseudo-likelihood method is described. The theoretical materials are evaluated in simulation studies and the application of the cosine model in a real example is presented.

Key words: Circular densities, Bivariate Von-Mises distribution, Cosine model, Pseudo-likelihood estimator.

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