مجله علوم آماری، بهار و تابستان ۱۳۹۳
جلد ۸، شماره ۱، ص ۱۱۱-۱۲۷

ویژگی‌های توزیع آماری برای زوایای دوسطحی

آناهیتا نودهی، موسی گل‌علی‌زاده
گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۲/۸/۱۹ تاریخ آخرین بازنگری: ۱۳۹۳/۴/۲۵

چکیده : مدل کسینوسی توزیع ون میسز دو متغیره، که تا حدودی رفتاری مشابه توزیع نرمال دومتغیره دارد، برای نمایش تغییرات احتمالاتی توام زوایای دوسطحی پیشنهاد شده است. از ویژگی‌های بارز این توزیع داشتن چگالی شرطی ون میسز یک متغیره است. اما توزیع حاشیه‌ای آن بسته به پارامترهای درگیر مسئله صورت‌های متفاوتی به خود می‌گیرد و به‌طور کلی شکل بسته‌ای ندارد. این موضوع استنباط آماری راجع به پارامترهای توزیع را با مشکلات خاصی همراه می‌کند. در مقالهٔ حاضر توزیع مورد اشاره و ویژگی‌های آماری آن مطالعه و سپس نحوه نمونه‌گیری از آن با الگوریتم رد و پذیرش تشریح می‌شود. به‌دلیل محدودیت دوره‌ای بودن توام زوایای دوسطحی، مشکلات مربوط به انتخاب توزیع‌های کاندید مناسب مطرح و از ویژگی‌های چگالی شرطی آن برای رفع این معضل بهره گرفته می‌شود.

واژه‌های کلیدی : آمار دایره‌ای، توزیع ون میسز، زوایای دوسطحی، چنبره، الگوریتم رد-پذیرش

* آدرس الکترونیک مسئول مقاله: موسی گل‌علی‌زاده، golalizadeh@modares.ac.ir
* کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲H۱۱، ۶۲P۱۰

۱ مقدمه

از جمله کاربردهای آمار در سایر علوم، استفاده از اطلاعات و منابع موجود برای مدل‌بندی و تحلیل نتایج است. فرض اساسی بسیاری از روش‌های آماری که برای مدل‌بندی داده‌ها و پیش‌بینی نتایج حاصل از آن مورد استفاده قرار می‌گیرد این است که داده‌های مورد مطالعه متعلق به فضای اُقلیدسی‌اند. توانایی بشر در عصر جدید او را قادر به اخذ داده‌هایی کرده است که ماهیتی نااُقلیدسی دارند. جهت‌ها مثالی از داده‌های نااُقلیدسی هستند. یک مجموعه از چنین مشاهداتی که با لحاظ نمودن جهت‌شان مشخص می‌شوند را داده‌های سویی[۱] می‌نامند. ویژگی اصلی این‌گونه داده‌ها داشتن دور یا جهت خاص است. مشاهداتی که از این نوع متغیرها ناشی می‌شوند در علومی مانند زیست‌شناسی، زمین‌شناسی، جغرافیا و ... به وفور دیده می‌شوند (ماردیا، ۱۹۷۵). به‌نظر می‌رسد به‌کارگیری نتایج مفاهیم موجود در فضای اُقلیدسی برای این داده‌ها نتایج صحیحی به‌همراه نداشته باشد.

یکی از زیر شاخه‌های مهم آمار سویی آمار دایره‌ای[۲] است که در چند دههٔ اخیر بسط و گسترش فراوانی یافته است (ماردیا و جاپ،۲۰۰۰). اگرچه مسائل مربوط به جهت‌های یک بُعدی تا حد زیادی آسان و سرراست هستند، اما موضوعات آماری راجع به جهت‌های دوبُعدی از پیچیده‌گی‌های خاصی برخوردار است. اندازه‌گیری حاصل از جهت‌های دوبُعدی معمولاً به‌صورت زوایا نشان داده می‌شوند. به‌طور کلی، اندازه‌گیری جهت‌ها بر اساس جهت صفر[۳] (نقطه‌ی شروع) و حالت دوران[۴] (موافق یا مخالف جهت حرکت عقربه‌های ساعت) مشخص می‌شود. با این حال، باید توجه داشت که نتایج تحلیل‌های آماری شامل خلاصه‌سازی و استنباط نباید وابسته به مقادیر دلخواه مبدا و حالت دوران باشند. به زبان ساده، باید محاسبات آماری طوری تعدیل شوند تا متاثر از پارامترهای مزاحم این‌گونه داده‌ها شامل نقطه‌ی مبدا، وضعیت دوران و ویژگی‌های دوره‌ای زوایا نباشد. این نکته و جنبه‌های متفاوت آماری دیگر مربوط به جهت‌ها در منابع استاندارد مثل ماردیا (۱۹۷۲)، ماردیا و جاپ (۲۰۰۰) و جامالاماداکا و سن گوپتا (۲۰۰۱) مورد مطالعه قرار گرفته است.

وضعیت‌هایی متفاوت از پدیدهای تصادفی مربوط به آمار دایره‌ای وجود دارد که داده‌های حاصل دو یا چند متغیره هستند. به‌عنوان مثال، چنین حالتی در مطالعه جفت زوایای دو

---

۱ Directional Data

۲ Circular data

۳ Zero direction

۴ Rotation

سطحی[۵] برای چند پروتئین اتفاق می‌افتد. می‌توان حدس زد که تعمیم نتایج حاصل از آمار دایره‌ای تک متغیره و هم‌چنین لحاظ فضای توپولوژیکی که جفت زوایا در آنجا قرار می‌گیرند نیازمند توجه خاصی است. به‌ویژه، ماردیا و همکاران (۲۰۰۷) نشان دادند زوایای دو سطحی بر روی یک چنبره قرار دارند. اخیراً تعیین ساختار دینامیکی و کارکرد مولکول‌های زیستی در فضای میکروسکوپی باعث گشودن زمینهٔ تحقیقاتی جدیدی برای محققین آماری علاقه‌مند به ویژگی‌های هندسی مثلاً با رویکرد احتمالات هندسی شده است طوری که هدف اصلی این فعالیت‌ها تعیین ساختار پروتئین است (درایدن و همکاران، ۲۰۰۵). نکته قابل تأمل در این حوزه که نقش مطالعات آماری را پررنگ‌تر می‌کند این است که می‌توان با دراختیار داشتن زوایای دوسطحی بین زوج اتم‌های متوالی یک پروتئین ساختار هندسی آن را به‌طور منحصربفرد به‌دست آورد (الماسی و بری ، ۱۹۹۹). بنابراین شناخت بیشتر توزیع احتمالاتی این زوایا و نحوهٔ نمونه‌گیری از آن می‌تواند برای نیل به حل پاره‌ای از مشکلات موجود در بیوانفورماتیک به‌خصوص تعیین ساختار پروتئین‌ها مد نظر قرار گیرد.

فعالیت‌های علمی برای معرفی و مطالعه توزیع دایره‌ای دومتغیره توسط ماردیا (۱۹۷۵) شروع شد که آنرا ون میسز دو متغیره نامید. او بیان داشت که چنین توزیعی بسیار شبیه توزیع نرمال دومتغیره‌ای است که بر روی چنبره قرار می‌گیرد. یکی از نگرانی‌ها در مورد آن توزیع وجود ۸ پارامتر بود در حالی‌که انتظار می‌رفت تنها ۵ پارامتر (بنا به ماهیت توزیع نرمال دومتغیره) برای این امر نیاز باشد. برای رفع این مشکل، ریوست (۱۹۸۸) همان توزیع را طوری تعدیل نمود که تعداد پارامترهای مدل به ۶ کاهش یافت. برای استفاده از این توزیع ۶ پارامتری در مثال‌های کاربردی مربوط به پروتئین‌ها، سینگ و همکاران (۲۰۰۲) سه زیر مدل به نام‌های سینوس، کسینوس با اثرات متقابل مثبت و منفی را معرفی کردند. اما جزئیات بیشتر در مورد دو زیر مدل آخر بترتیب توسط ماردیا و همکاران (۲۰۰۷) و کنت و همکاران (۲۰۰۸) تشریح شد. در هر دو فعالیت بطور گذرا به نحوه اخذ نمونه‌های تصادفی از توزیع‌های مربوطه اشاره شد ولی جزئیات الگوریتم مناسب و مشکلات مبتلا به آن مدنظر قرار نگرفت. بیان این نکته ضروری است که اخیرا، ماردیا (۲۰۱۰) با رویکرد بیزی توزیع ون میسز دو متغیره را مورد مطالعه قرار داد.

به دلیل اهمیت توزیع ون میسز دو متغیره در مدل‌بندی جفت زاویای دوسطحی و به‌ویژه تعیین ساختارهای متفاوت زنجیره اصلی پروتئین‌ها، که از مسائل پر اهمیت علوم زیستی است، توجه این مقاله معطوف به این توزیع شده است. به‌طور دقیق‌تر، نحوه اخذ نمونه‌های

---

۵ Dihedral angles

تصادفی از آن‌ها (همراه با جزئیات مورد نیاز) تشریح می‌شود. برای این منظور توزیع‌های حاشیه‌ای و شرطی‌های مورد نیاز محاسبه و در الگوریتم رد و پذیرش مورد استفاده قرار می‌گیرد. نتایجی از شبیه‌سازی‌های صورت گرفته نمایش داده خواهد شد. نکته حائز اهمیت این است که با شناخت توزیع‌های احتمالاتی مرتبط با توزیع ون میسز دو متغیره و به‌ویژه اخذ مشاهداتی از آن می‌توان به استنباط آماری راجع به پارامترهای توزیع پرداخت.

در بخش ۲ خلاصه‌ای از آمار دایره‌ای و توزیع ون میسز تشریح می‌شود. سپس، توزیع‌های آماری مناسب بر روی چنبره و ویژگی‌های آماری مهم آن‌ها در بخش ۳ ارائه می‌شود. بخش ۴ دربرگیرنده نحوه شبیه‌سازی از مدل کسینوسی توزیع ون میسز دو متغیره و نمایش هندسی نمونه‌های تولید شده است. مقاله با بحث و نتیجه‌گیری کلی خاتمه می‌یابد.

۲ آمار دایره‌ای و توزیع‌های روی آن

معمولاً داده‌های زاویه‌ای به‌صورت زوایا یا نقاط بر روی محیط دایره‌ی واحد نمایش داده می‌شوند. به‌طور کلی، موقعیت جهت‌ها را می‌توان به‌صورت یکتا از طریق مختصات دوبُعدی تعیین کرد. برای این منظور می‌توان از مختصات دکارتی استفاده نمود که به این ترتیب، هر نقطه مانند p به‌صورت (x, y) که در مختصات قطبی به‌صورت (r, α) که در آن r فاصله نقاط تا نقطهٔ مرکز o و α جهت آن‌هاست، نمایش داده می‌شود. مختصات قطبی و دکارتی با استفاده از توابع مثلثاتی سینوس و کسینوس به‌صورت $x = r\cos\alpha$ و $y = r\sin\alpha$ قابل تبدیل به یکدیگر هستند. از آنجایی‌که در تحلیل زاویه‌ای محقق علاقه‌مند به بررسی جهت‌ها به‌جای مقادیر بردارها است و طبق قرارداد این بردارها دارای اندازهٔ واحد هستند پس هرجهت متناظر با یک نقطه بر روی محیط دایره واحد است. به زبان ریاضی می‌توان نوشت:

$$(۱, \alpha) \Leftrightarrow (x = \cos\alpha, y = \sin\alpha).$$

در واقع دیدگاه کلی در آمار دایره‌ای مبتنی بر زاویه α براساس بردار متناظرش یعنی (x, y) بر روی دایرهٔ واحد است. برخلاف دامنهٔ دوره‌ای در α، بردارهای (x, y) در فضای خطی و میانگین و فاصله‌ها همانند فضای اُقلیدسی تعریف می‌شوند. به‌علاوه، استفاده از بردارها باعث سادگی محاسبات می‌شود و در پی آن به‌کارگیری بسیاری از ابزارهای آمار خطی[۶] میسر می‌گردد.

[۶] Linear statistics

برای متغیر تصادفی دایره‌ای X، با فضای نمونه S، تابع توزیع $F_X(x)$ علاوه بر خواص مرسوم بایستی دارای ویژگی‌های زیر باشد:

$$F_X(x) = Pr(\circ < \theta \leq x), \quad \circ \leq x \leq \mathbf{2}\pi,$$

$$F_X(x + \mathbf{2}\pi) - F_X(x) = \mathbf{1}, \quad -\infty < x < +\infty$$

رابطهٔ آخر بیان می‌کند که احتمال طول کمان $\mathbf{2}\pi$ برابر یک است. برای هر تابع توزیع متغیر دایره‌ای، تابع چگالی متناظر با آن به کمک روابط آمار خطی به‌دست می‌آید. در حالت پیوسته تابع چگالی احتمال به‌ازای هر عدد صحیح k در تساوی $f(\theta) = f(\theta + \mathbf{2}k\pi)$ صدق می‌کند (جامالاماداکا و سن گوپتا،۲۰۰۱).

توزیع ون میسز یک توزیع دایره‌ای است که نقشی مشابه توزیع نرمال در فضای اُقلیدسی دارد (ماردیا، ۱۹۷۵). تا قبل از کشف ایزوتوپ‌ها در اتم تصور می‌شد که وزن اتمی عناصر عددی صحیح است. با در نظر گرفتن بخش اعشاری وزن ۲۴ عنصر سبک ریچارد ون میسز (۱۹۱۸) نشان داد که قسمت کسری مربوطه می‌تواند به‌عنوان متغیر تصادفی بر روی دایره تلقی شود. ایشان با نمایش هندسی و بررسی علمی آن سبب ایجاد تحول شگرفی در فیزیک اتمی گشت. به همین دلیل توزیع پیشنهادی ایشان به افتخار وی به توزیع ون میسز معروف شد. البته، بنا به اعتقاد جامالاماداکا و سن گوپتا (۲۰۰۱) قبل از ون میسز پل لانجوین (۱۹۰۵) از این توزیع در برخی مباحث فیزیکی استفاده کرده بود. گذر تاریخی این توزیع به همین‌جا ختم نمی‌شود. بلکه در پی سالیان متمادی با بسط آن در بعدهای بالاتر اهمیت آن بیش از پیش مشخص شد (ماردیا و جاب، ۲۰۰۰).

اگر θ دارای توزیع ون میسز با پارامترهای μ و κ باشد که با نماد $\theta \sim VM(\mu, \kappa)$ نمایش داده می‌شود، آنگاه تابع چگالی احتمال آن به‌صورت

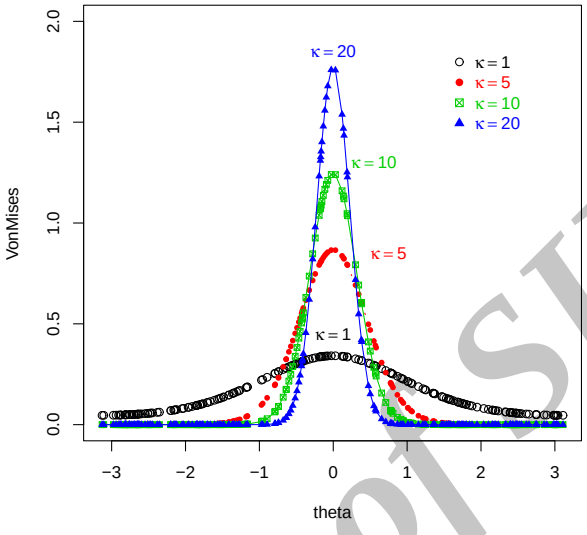
$$f(\theta) = \frac{\mathbf{1}}{\mathbf{2}\pi I_\circ(\kappa)} e^{\kappa \cos(\theta - \mu)} \quad , \quad -\pi \leq \theta \leq \pi, \qquad (\mathbf{1})$$

نوشته می‌شود، که در آن پارامتر μ میانگین توزیع، κ پارامتر تمرکز[۷] و $I_\circ(\kappa)$ تابع بسِل از مرتبهٔ صفر به‌صورت

$$I_\circ(k) = \frac{\mathbf{1}}{\mathbf{2}\pi} \int_\circ^{\mathbf{2}\pi} e^{\kappa \cos(\theta)} d\theta = \sum_{r=\circ}^{\infty} (\frac{\kappa}{\mathbf{2}})^{\mathbf{2}r} (\frac{\mathbf{1}}{r!})^{\mathbf{2}}$$

است (آبراموویچ و اِستگان ، ۱۹۶۵). توزیع ون میسز، متقارن و تک مدی در نقطهٔ میانگین μ است. هنگامی که $\kappa = \circ$ توزیع ون میسز به توزیع یکنواخت تبدیل خواهد شد. اما اگر κ

[۷] Concentration parameter



شکل ۱: رفتار توزیع ون میسز به‌ازای مقادیر متفاوت پارامتر تمرکز

افزایش یابد، جرم توزیع ون میسز حوالی μ متراکم شده و به همین دلیل به κ پارامتر تمرکز گفته می‌شود. شکل ۱ رفتار این توزیع را به‌ازای مقادیر متفاوت پارامتر تمرکز نشان می‌دهد.

۳ چنبره و توزیع‌های آماری بر روی آن

در هندسه، چنبره یکی از اشکال سه بُعدی است که از چرخاندن یک دایره در فضای سه‌بُعدی حول محوری هم‌صفحه ولی غیرمماس با آن دایره به وجود می‌آید. از نقطه‌نظر توپولوژی، چنبره همانند دستگاه مختصات دکارتی است که از ضرب دو دایره به‌دست می‌آید یعنی اگر S^1 نمایانگر دایره و T^2 نمایانگر چنبره باشد آنگاه $T^2 = S^1 \times S^1$. به‌علاوه، چنبره را می‌توان به‌صورت پارامتری توسط روابط

$$X = (R + r\cos\theta_1)\cos\theta_2, \qquad Y = (R + r\cos\theta_1)\sin\theta_2, \qquad Z = r\sin\theta_1.$$

نوشت، که در آن $(\theta_1, \theta_2) \in [-\pi, \pi]$ و باتوجه به اینکه چنبره از ضرب دو دایره به‌دست آمده است، R شعاع دایرهٔ بزرگتر و r شعاع دایرهٔ کوچکتر در چنبره است (هاچِر، ۲۰۰۲).

در دستگاه مختصات دکارتی معادلهٔ چنبره به‌صورت

$$(\sqrt{X^2+Y^2}-R)^2+Z^2=r^2.$$

است که آن را می‌توان به‌طور هم‌ارز به‌صورت

$$(X^2+Y^2+Z^2+R^2-r^2)^2=4R^2(X^2+Y^2)$$

نوشت. لازم به‌ذکر است که در تحلیل داده‌های نااُقلیدسی فضای توپولوژیکی قرارگیری داده‌ها مبنای استنباط و مدل‌بندی آماری است و از آنجایی‌که زوایای دو سطحی بر روی چنبره استاندارد ($R=2, r=1$) قرار دارند (ماردیا و همکاران، ۲۰۰۷)، تمام تحلیل‌ها و استنباط آماری این زوایا بر روی این حالت خاص از چنبره انجام گرفته است.

فرض کنید ϑ_1, ϑ_2 دو متغیر تصادفی زاویه‌ای و θ_1 و θ_2 مقادیر مشاهده شدهٔ آن‌ها هستند. توجه کنید که در سراسر این بخش تکیه‌گاه چنین متغیرهای زاویه‌ای $[-\pi, \pi]$ است. ماردیا (۱۹۷۵) تابع چگالی احتمال توام آن‌ها را به‌صورت

$$\begin{aligned} f(\theta_1,\theta_2) &= \exp\big\{[\cos(\theta_1-\mu_1), \sin(\theta_1-\mu_1)]A[\cos(\theta_2-\mu_2), \sin(\theta_2-\mu_2)]^T \\ &+ \kappa_1\cos(\theta_1-\mu_1)+\kappa_2\cos(\theta_2-\mu_2)\big\}\{C(\cdot)\}^{-1}, \end{aligned} \qquad (2)$$

پیشنهاد داد که در آن $\kappa_1, \kappa_2 \geq 0$، مقادیری ثابت و معیارهایی از تمرکز توزیع، پارامترهای $\mu_1, \mu_2 \in [-\pi, \pi]$، معرف میانگین توزیع، $C(\cdot) = C(\kappa_1, \kappa_2, A)$ ثابت نرمال ساز و ماتریس $A = [a_{ij}]$ یک ماتریس 2×2 و معیاری از پراکندگی توزیع هستند. همان‌گونه که از رابطهٔ (۲) ملاحظه می‌شود چگالی حاصل ۸ پارامتر دارد. ریوست (۱۹۸۸) زیر مدلی از این توزیع را به‌صورت

$$\begin{aligned} f(\theta_1,\theta_2) &= \exp\{\alpha\cos(\theta_1-\mu_1)\cos(\theta_2-\mu_2)+\beta\sin(\theta_1-\mu_1)\sin(\theta_2-\mu_2) \\ &+ \kappa_1\cos(\theta_1-\mu_1)+\kappa_2\cos(\theta_2-\mu_2)\}\{C(\cdot)\}^{-1}, \end{aligned} \qquad (3)$$

معرفی کرد که دارای ۶ پارامتر است. سالیان بعد، سینگ و همکاران (۲۰۰۲) با الگوبرداری از توزیع نرمال دومتغیره حالت‌های خاصی از این زیرمدل را مورد مطالعه قرار دادند و آن‌ها را در چند مثال کاربردی به‌کار بردند. سه زیر مدل پیشنهادی توسط آن‌ها به‌نام‌های مدل سینوسی، کسینوسی با اثرات متقابل مثبت و مدل کسینوسی با اثرات متقابل منفی معروف است.

اگر مولفه‌های ماتریس $A = [a_{ij}]$ در معادله (۲) با مقادیر $a_{11} = a_{12} = a_{21} = 0$ و $a_{22} = \eta$ جایگزین شود، آنگاه مدل سینوسی دارای تابع چگالی به‌صورت

$$f(\theta_1, \theta_2) = \frac{e^{\kappa_1 \cos(\theta_1 - \mu_1) + \kappa_2 \cos(\theta_2 - \mu_2) + \eta \sin(\theta_1 - \mu_1) \sin(\theta_2 - \mu_2)}}{C(\kappa_1, \kappa_2, \eta)} \qquad (۴)$$

خواهد بود، که در آن $-\infty < \eta < +\infty$ معیاری از وابستگی بین ϑ_1 و ϑ_2 است. در این حالت اگر $\eta = 0$ آنگاه ϑ_1 و ϑ_2 مستقل بوده و هرکدام دارای توزیع ون میسز تک متغیره خواهند بود. توجه کنید که با قرار دادن $\alpha = 0, \beta = \eta$ در رابطهٔ (۳) نیز مدل سینوسی (۴) به‌دست خواهد آمد. سینگ و همکاران (۲۰۰۲) مقدار ثابت نرمال ساز در مدل سینوس را به‌صورت

$$C(\kappa_1, \kappa_2, \eta) = \{4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} (\frac{\eta}{2})^{2m} \kappa_1^{-m} I_m(\kappa_1) \kappa_2^{-m} I_m(\kappa_2)\}^{-1}$$

به‌دست آوردند، که در آن $I_r(\kappa)$ تابع بسل نوع اول از مرتبهٔ r است. آنها همچنین نشان دادند که توزیع‌های شرطی از توزیع (۴)، وَن میسز و توزیع‌های حاشیه‌ای، متقارن در میانگین دایره‌ای هستند. توزیع‌های حاشیه‌ای هم می‌توانند تک مدی و هم دو مدی باشند و با ایجاد شرایطی در پارامترها توزیع شکل خاصی به خود می‌گیرد. در توزیع‌های تک مدی، هنگامی که پارامترهای تمرکز کوچک باشند، حاشیه‌ها تقریباً از توزیع ون میسز پیروی خواهند کرد.

ماردیا و همکاران (۲۰۰۷) با قرار دادن $\alpha = \beta = -\kappa_3$ در رابطهٔ (۳) مدل کسینوس با اثرات متقابل مثبت با تابع چگالی به‌صورت

$$f_{\cos PI}(\theta_1, \theta_2) \propto e^{\kappa_1 \cos(\theta_1 - \mu_1) + \kappa_2 \cos(\theta_2 - \mu_2) - \kappa_3 \cos(\theta_1 - \mu_1 - \theta_2 + \mu_2)}.$$

معرفی کردند. کنت و همکاران (۲۰۰۸) تابع چگالی دیگری پیشنهاد کردند که مدل کسینوسی با اثرات متقابل منفی را به نمایش می‌گذارد. آن تابع چگالی به‌صورت

$$f_{\cos NI}(\theta_1, \theta_2) \propto e^{\kappa_1 \cos(\theta_1 - \mu_1) + \kappa_2 \cos(\theta_2 - \mu_2) - \kappa_3 \cos(\theta_1 - \mu_1 + \theta_2 - \mu_2)}$$

است، که در آن $\kappa_1 > \kappa_3 > 0$ و $\kappa_2 > \kappa_3 > 0$. این توزیع هنگامی که نوسانات در متغیر دایره‌ای کوچک باشد، تقریباً نرمال دومتغیره است. توجه کنید که تابع چگالی مدل کسینوس با اثرات متقابل منفی با قرار دادن $\alpha = -\beta = -\kappa_3$ در رابطهٔ (۳) به‌دست می‌آید. ماردیا (۲۰۰۹) مقدار ثابت نرمال ساز در این مدل را به‌صورت

$$C(\kappa_1, \kappa_2, \kappa_3) = \{4\pi^2 [I_0(\kappa_1) I_0(\kappa_2) I_0(\kappa_3) + 2 \sum_{p=0}^{\infty} I_p(\kappa_1) I_p(\kappa_2) I_p(\kappa_3)]\}^{-1}.$$

پیشنهاد کرد. برخی از ویژگی‌های مدل‌های سینوسی و کسینوسی عبارتند از:

الف) هر دو مدل سینوسی و کسینوسی متقارن‌اند، یعنی $f(\theta_1,\theta_2)=f(-\theta_1,-\theta_2)$.

ب) در بیشتر مواقع، اختلاف اندکی مابین مدل‌های سینوسی و کسینوسی وجود دارد. به‌علاوه، هنگامی‌که پارامتر تمرکز زیاد باشد، هر دو مدل معادل برازش توزیع نرمال دومتغیره در صفحهٔ مماسی در نقطهٔ میانگین (μ_1,μ_2) است.

ج) هنگامی که $|\eta|$ در مدل سینوسی افزایش یابد، چگالی آن دو مدی خواهد شد در حالی که در مدل کسینوسی با افزایش مقادیر مثبت κ_3 این اتفاق خواهد افتاد (شکل ۲).

در شبیه‌سازی از توزیع دومتغیرهٔ ون میسز، توابع چگالی شرطی و حاشیه‌ای نقشی اساسی ایفا می‌کنند. با پیروی از ماردیا و همکاران (۲۰۰۷) و با فرض اینکه $\mu_1=\mu_2=0$، توابع چگالی احتمال مدل سینوسی و کسینوسی با اثرات متقابل مثبت و منفی، به‌ترتیب، به فرم

$$
\begin{aligned}
f_{\sin}(\theta_1,\theta_2)&=C(\kappa_1,\kappa_2,\eta)\exp\{\kappa_1\cos\theta_1+\kappa_2\cos\theta_2+\eta\sin\theta_1\sin\theta_2\}\\
f_{\cos PI}(\theta_1,\theta_2)&=C(\kappa_1,\kappa_2,\kappa_3)\exp\{\kappa_1\cos\theta_1+\kappa_2\cos\theta_2-\kappa_3\cos(\theta_1-\theta_2)\}\qquad(5)\\
f_{\cos NI}(\theta_1,\theta_2)&=C(\kappa_1,\kappa_2,\kappa_3)\exp\{\kappa_1\cos\theta_1+\kappa_2\cos\theta_2-\kappa_3\cos(\theta_1+\theta_2)\}
\end{aligned}
$$

هستند. برای محاسبهٔ توزیع‌های حاشیه‌ای و شرطی، پارامترهای جدید $\alpha(\theta_1)$ و $\beta(\theta_1)$ که به اختصار با α و β نمایش داده طوری تعریف می‌شود که در روابط $\kappa_2=\alpha\cos\beta$ و $\eta\sin\theta_1=\alpha\sin\beta$ صدق کنند. بنابراین

$$\alpha=(\kappa_2^2+\eta^2\sin^2\theta_1)^{\frac{1}{2}},\quad \beta=\arctan(\frac{\eta}{\kappa_2}\sin\theta_1).$$

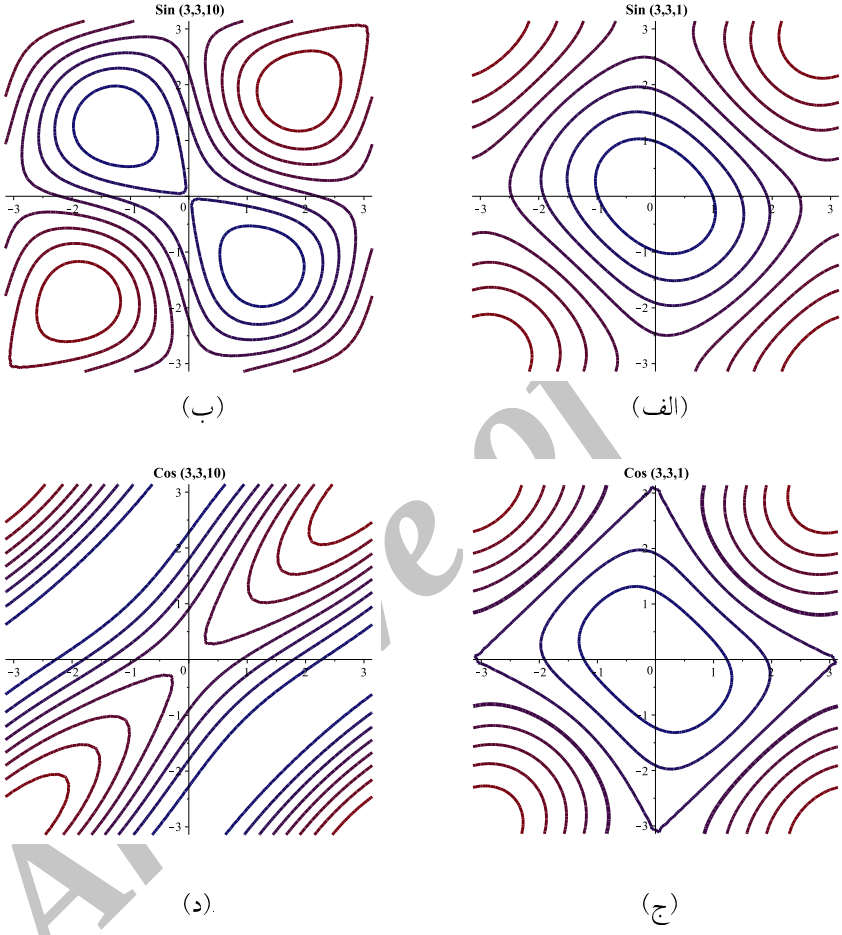
حال برای محاسبهٔ تابع چگالی احتمال حاشیه‌ای ϑ_1 در مدل سینوسی داریم:

$$
\begin{aligned}
f_{\sin}(\theta_1) &= \int_{-\pi}^{\pi} f_{\sin}(\theta_1,\theta_2)d\theta_2\\
&= C(\kappa_1,\kappa_2,\eta)^{-1}2\pi I_0(\alpha(\theta_1))\exp\left\{\kappa_1\cos\theta_1\right\},
\end{aligned}
$$

که در آن $I_0(\cdot)$ تابع بِسِل از مرتبهٔ صفر است. ملاحظه می‌شود این تابع چگالی شباهتی با توزیع‌های معمول بر روی دایره ندارد (ماردیا و جاپ، ۲۰۰۰).

با توجه به محاسبه توزیع‌های حاشیه‌ای و تئوری احتمالات، محاسبه توزیع‌های شرطی ساده خواهد بود. به زبانی دقیق‌تر، تابع چگالی شرطی ϑ_2 به شرط $\vartheta_1=\theta_1$ برابر

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



(الف) (ب)

(ج) (د)

شکل ۲: نمودار تراز لگاریتم تابع چگالی مدل سینوس با مقادیر پارامتر (الف): (۳, ۳, ۱) و (ب): (۳, ۳, ۱۰). نمودار تراز لگاریتم تابع چگالی مدل کسینوس با مقادیر پارامتر (ج): (۳, ۳, ۱) و (د): (۳, ۳, ۱۰).

خواهد شد. می‌توان ملاحظه کرد که تابع چگالی شرطی $\vartheta_۲$ به شرط $\vartheta_۱ = \theta_۱$ دارای توزیع وَن میسز با پارامتر تمرکز $\alpha(\theta_۱)$ و میانگین زاویه‌ای $\beta(\theta_۱)$ است.

برای محاسبهٔ توابع چگالی حاشیه‌ای و شرطی مدل کسینوسی پارامترهای جدید $\alpha(\theta_۲)$ و $\beta(\theta_۲)$ به‌صورت

$$\begin{aligned} \alpha(\theta_۲) &= \sqrt{(\kappa_۱^۲ + \kappa_۳^۲ - ۲\kappa_۱\kappa_۳\cos\theta_۲)}, \\ \beta(\theta_۲) &= \arctan\left(\frac{-\kappa_۳\sin\theta_۲}{\kappa_۱ - \kappa_۳\cos\theta_۲}\right) \end{aligned}$$

تعریف می‌شود. با بررسی توابع حاشیه‌ای و شرطی مدل کسینوسی با اثرات متقابل مثبت داریم:

$$\begin{aligned} f_{\cos}(\theta_۲) &= \int_{-\pi}^{\pi} f_{\cos}(\theta_۱, \theta_۲)\, d\theta_۱ \\ &= C(\kappa_۱, \kappa_۲, \kappa_۳)^{-۱} ۲\pi I_۰(\alpha(\theta_۲)) \exp\{\kappa_۲\cos\theta_۲\}. \end{aligned} \qquad (۶)$$

اکنون با توجه به توزیع توام مدل کسینوسی و توزیع حاشیه‌ای حاصل در (۶) توزیع چگالی شرطی مدل کسینوسی به آسانی به‌صورت

$$\begin{aligned} f_{\cos}(\theta_۱|\theta_۲) &= \frac{f_{\cos}(\theta_۱, \theta_۲)}{f_{\cos}(\theta_۲)} \\ &= \frac{۱}{۲\pi I_۰(\alpha(\theta_۲))} \exp\{\alpha(\theta_۲)\cos(\theta_۱ - \beta(\theta_۲))\} \end{aligned}$$

قابل محاسبه است. همان طور که ملاحظه می‌شود توزیع وَن میسز با پارامتر تمرکز $\alpha(\theta_۲)$ و میانگین زاویه‌ای $\beta(\theta_۲)$ خواهد بود.

۴ شبیه‌سازی از مدل کسینوسی توزیع ون میسز دومتغیره

در علوم زیستی و به‌ویژه در بیوانفورماتیک تولید پروتئین‌ها با ویژگی‌های متفاوت بسیار پرهزینه است. از نقطه‌نظر آماری نیز در خیلی از مواقع دسترسی به هر حجمی از داده‌های واقعی غیرممکن بنظر می‌رسد (اِشمیت و تیلور، ۱۹۷۰). در عوض انجام شبیه‌سازی از مدل‌ها یا ساختارهای تصادفی کمک شایانی در این موضوع می‌کند. به‌علاوه، معمولا با شبیه‌سازی می‌توان به راحتی و با صرف هزینه کم به حجم قابل توجهی از داده‌ها دست یافت. در نهایت از شبیه‌سازی داده‌ها می‌توان برای مدل‌بندی و یا پیش‌بینی اینکه چگونه تغییرات حجم داده‌ها بر روی نتایج حاصل از استنباط آماری تأثیر می‌گذارد، استفاده کرد. لذا، در این بخش نمونه‌های تصادفی که نقش مشاهدات زاویه‌ای بر روی چنبره را بازی می‌کنند، تولید می‌شود.

با نگاه به توزیع کسینوسی با اثرات متقابل مثبت معرفی شده در بخش ۳ ملاحظه می‌شود که چگالی مربوطه فرم بسته‌ای ندارد و لذا شبیه‌سازی مستقیم از آن آسان نخواهد بود. با این‌حال چون چگالی شرطی ϑ_1 به‌شرط $\vartheta_2 = \theta_2$ توزیع ون میسز است که شبیه‌سازی از آن به‌وسیلهٔ کتابخانه Circular در نرم‌افزار آماری R میسر خواهد بود، می‌توان شبیه‌سازی از این چگالی شرطی را با شبیه‌سازی از توزیع حاشیه‌ای ϑ_2 درهم آمیخت و زوج نمونه‌های $(\vartheta_1, \vartheta_2)$ را به‌دست آورد.

با دقت در توزیع حاشیه‌ای ϑ_2 در توزیع وَن میسز دو متغیره مدل کسینوسی درمی‌یابیم شبیه‌سازی مستقیم از آن نیز میسر نیست. اما می‌توان از الگوریتم‌های مختلفی از جمله الگوریتم رد و پذیرش، برای تولید نمونه تصادفی از این توزیع حاشیه‌ای استفاده نمود. برای این منظور باید توجه شود که توزیع کاندید را باید طوری درنظر گرفت که هم‌رفتار و دارای تکیه‌گاه یکسان با توزیع موردنظر باشد (رابرت و کسلا،۲۰۰۴). برای تعیین توزیع نامزد با دقت در فرم توزیع حاشیه‌ای ϑ_2 در توزیع وَن میسز دو متغیره مدل کسینوسی و مقایسهٔ آن با رابطهٔ (۱) ملاحظه می‌شود که قسمت نمایی آن‌ها بسیار به هم شبیه بوده بنابراین میانگین توزیع را صفر و پارامتر تمرکز (κ_2) را طوری در نظر گرفته که توزیع نامزد، توزیع حاشیه‌ای را در سراسر دامنه تغییرات θ_2 پوشش[۸] دهد.

گام‌های الگوریتم رد و پذیرش پشنهادی برای تولید نمونه از توزیع کسینوسی با اثرات متقابل مثبت به‌صورت زیر است:

گام ۱ : از توزیع کاندید $h(\theta_2) = VM(۰, ۰/۵)$ داده تولید کنید.

گام ۲ : ثابت L طوری در نظر گرفته شود که به‌ازای هر θ_2 در شرایط $f(\theta_2) < L \times h(\theta_2)$ ، $L = \max\left\{\frac{f(\theta_2)}{h(\theta_2)}, ۱\right\}$ صدق کند.

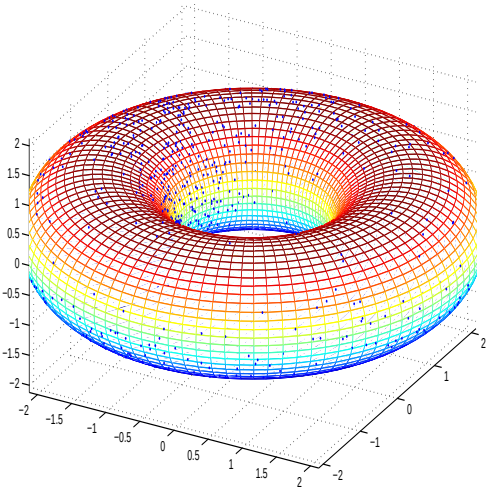
گام ۳ : از توزیع $U(۰, ۱)$ داده تولید کنید.

گام ۴ : اگر شرط $U < \frac{f(\theta_2)}{L \times h(\theta_2)}$ برقرار باشد، مقدار θ_2 به‌عنوان مقدار تصادفی از توزیع موردنظر $f(\theta_2)$ پذیرفته شود و در غیر این صورت به **گام ۱** برگردید.

با داشتن θ_2، برای تولید متغیر تصادفی $\vartheta_1 = \theta_1$، از توزیع شرطی $f(\theta_1|\theta_2)$ که دارای توزیع ون میسز با پارامترهای $\{\mu = \beta(\theta_2), \kappa = \alpha(\theta_2)\}$ است، با استفاده از بستهٔ آماری Circular (در نرم‌افزار R) میسر است.

---

[۸] Dominate



شکل ۳: نمودار داده‌های شبیه‌سازی شده مدل کسینوسی از توزیع ون میسز دومتغیره بر روی چنبره

برای مدل کسینوسی با اثرات متقابل مثبت پارامترهای توزیع به‌صورت $\kappa_1 = ۱/۵, \kappa_2 = ۱/۷$ و $\kappa_3 = ۱/۳$ درنظر گرفته شد و نمونه‌ای به حجم ۱۰۰۰ از این توزیع تولید شده و فرم هندسی آن در شکل ۳ نشان داده شده است.

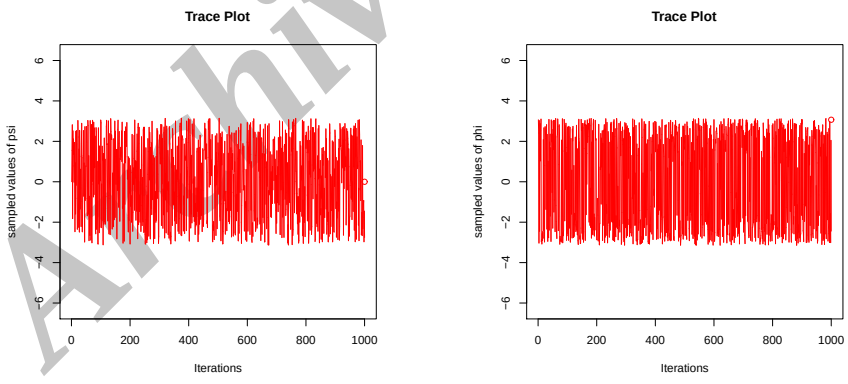
یکی از جنبه‌های مورد توجه نمونه‌های تولید شده بررسی دامنهٔ تغییرات آن‌هاست به این مفهوم که آیا در بازهٔ خودشان حضور دارند یا خیر. علاوه بر این، بررسی روند تغییرات آن‌ها در طی گام‌های شبیه‌سازی روند آمیختن[۹] نمونه‌های تولید شده را نمایان خواهد ساخت. برای این منظور، در شکل ۴ – الف نمودار اثر نمونه‌های تولید شده از تابع چگالی حاشیه‌ای مدل کسینوس با استفاده از الگوریتم رد و پذیرش با توزیع کاندید ذکر شده در گام ۱ رسم شده است. قابل ذکر است که درصد پذیرش نمونه‌ها در کل شبیه‌سازی **۴۹** درصد بود. همچنین، در شکل ۴ – ب نمونه‌های تولید شده از توزیع شرطی θ_1 به شرط θ_2 با پارامترهای (μ, κ) نمایش داده شده است. یکی از موارد استفاده از این نمونه‌های شبیه‌سازی شده کاربردشان در مبحث استنباط آماری است. موضوع برآورد پارامترهای این توزیع از مباحث

[۹] Mixing

نوین تحقیق در آمار دایره‌ای است که هامِلریک و همکاران (۲۰۱۲) به بخشی از آن اشاره کردند. به‌ویژه، آن‌ها تاکید کردند که محاسبه برآوردهای ماکسیمم درستنمایی مستلزم حل تعداد زیادی روابط غیرخطی است در حالی که برآوردگرهای گشتاوری ساختاری ساده‌تر دارند. اما باید توجه داشت که اولی جواب‌های دقیق‌تری از دومی خواهد داد. محاسبه برآوردگرهای ماکسیمم درستنمایی امری بسیار دشوار بوده که ترجیح داده شد در این مقاله به آن پرداخته نشود. اما با ۱۰۰ بار تکرار فرایند شبیه‌سازی، برآوردهای تجربی گشتاوری پارامترها همراه با خطای برآورد آنها به‌صورت

$$\widehat{\kappa}_1 = ۱/۲۰۴(۰/۰۸۴), \widehat{\kappa}_2 = ۱/۶۲۷(۰/۰۶۴), \widehat{\kappa}_3 = ۱/۴۴(۰/۱۰۴)$$

به‌دست آمده است. همان‌طور که ملاحظه می‌شود محاسبه بازه‌های اطمینان تقریبی دربرگیرنده مقادیر اولیه درنظرگرفته شده برای پارامترها نیست. این موضوع به نوعی تاییدی بر عدم اعتبار برآوردهای گشتاوری، همان‌طوری که هامِلریک و همکاران (۲۰۱۲) اشاره کردند، است.



(الف) (ب)

شکل ۴: نمودار نمونه‌های تولید شده از الف: چگالی $f_{\cos}(\theta_2)$ با استفاده از الگوریتم رد و پذیرش و ب: توزیع ون میسز $(\beta(\theta_2), \alpha(\theta_2))$.

بحث و نتیجه‌گیری

در این مقاله ضمن معرفی داده‌های زاویه‌ای که ماهیت ناأقلیدسی داشته و کاربرد فراوانی در مسائل کاربردی دارند، به تحلیل آن‌ها پرداخته و روش شبیه‌سازی از این گونه داده‌ها توضیح داده شد. نادیده گرفتن خاصیت زاویه‌ای بودن چنین متغیرهایی و اعمال روش‌های آمار خطی بر روی آن‌ها منجر به نتایج نادرست می‌گردد. بنابراین با مدنظر قرار دادن فضای ناأقلیدسی این داده‌ها و تعمیم توزیع‌ها و روش‌های آمار خطی در این فضا، سعی در پاسخگویی به سوالات متداول به‌خصوص در حوزه آمار غیرخطی شده است.

ازنقطه‌نظر تئوری، پیشرفت‌های خوبی در حوزهٔ آمار دایره‌ای و آمار روی چنبره صورت گرفته است ولی بسته‌های نرم‌افزاری کمتری در این زمینه موجود است. در نرم افزار Matlab بسته‌ای برای شبیه‌سازی از توزیع‌های آماری روی دایره و درحالت تک متغیره وجود دارد و در نرم‌افزار R بسته‌های Circular و CircStat برای این منظور موجودند. ولی درحال حاضر و با بررسی‌های صورت گرفته توسط نویسندگان این مقاله ملاحظه شد که در حالت دو یا چندمتغیره بستهٔ آماری خاصی موجود نیست.

از نقطه نظر کاربردی، می‌توان از این زوایا در حوزهٔ بیوانفورماتیک بخصوص تعیین ساختار پروتئین‌ها استفاده نمود. اما باید توجه داشت که پیچیدگی‌های منحصر به‌فرد ساختار پروتئین‌ها تنها با کمک توزیع‌های مورد اشاره در این مقاله و منابع آن قابل حل نیستند. در عوض، مدل‌بندی جامعی که شامل علوم زیستی، فیزیکی، کامپیوتری و آماری هستند می‌تواند مسیری را برای پاسخ به بخشی از مشکلات مورد اشاره هموار نمایند. بومسا و همکاران (۲۰۰۸) و ماردیا (۲۰۱۰) به گوشه‌ای از این مسائل در بررسی تعیین ساختار پروتئین‌ها اشاره کرده‌اند. ماردیا و همکاران (۲۰۰۸) به بررسی حالت چند متغیره توزیع ون میسز پرداخته و به کاربرد آن را در بیوانفورماتیک پرداخته‌اند. اخیرا، این حوزه از علم در علوم دیگر بسیار مطرح شد. اما پرداختن به این موضوع کمتر در منابع علمی آماری به‌چشم می‌خورد. امید است مقاله حاضر انگیزه‌ای برای محققین آمار غیرخطی که مایل به مطالعهٔ داده‌های زاویه‌ای و استنباط آماری آن‌ها درحالت دومتغیره‌اند، فراهم کرده باشد.

تقدیر و تشکر

نویسندگان از مسئولان مجله علوم آماری در فراهم کردن زمینه مشارکت علمی برای موضوعات جدید آماری تقدیر و از داوران محترم برای ارائه پیشنهادات مفید در بهبود این مقاله سپاسگزاری می‌نمایند.

مراجع

Abramowitz, M. and Stegun, I. A. (1965), *Handbook of Mathematical Functions*, Dover Publications, New York.

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008), A Generative, Probabilistic Model of Local Protein Structure, *Proceedings of the National Academy of Sciences of USA*, **105**, 8932-8937.

Dryden, I. L., Mian, S., Browne, W. J., Handley, K., di Nisio, R. and Rees, R. (2005), Statistical Analysis of SELDI Protein Chip Data from Breast Cancer Cell Lines Exposed to Chemotherapeutic Agents, In *S. Barber, P.D. Baxter, K.V. Mardia and R.E. Walls (Ed.), LASR 2005 - Quantitative Biology, Shape Analysis, and Wavelets*, University of Leeds, UK.

Elmaci, N. and Berry, R. S. (1999), Principal Coordinate Analysis on a Protein Model, *Journal of Chemistry-Physics*, **110**, 10606-10622.

Hamelryck, T., Mardia, K. V. and Ferkinghoff-Borg, J. (2012), *Bayesian Methods in Structural Bioinformatics*, Springer, Berlin.

Hatcher, A. (2002), *Algebraic Topology*, Cambridge University Press, Cambridge.

Jammalamadaka, S. R. and Sengupta, A. (2001), *Topics in Circular Statistics*, World Scientific Publishing Co., Inc, Singapore.

Kent, J. T., Mardia, K. V. and Taylor, C. C. (2008), Modelling Strategies for Bivariate Circular Data, In Gusnanto A., Barber, S., Baxter, P. D. and Mardia, K. V. editors, The Art and Science of Statistical Bioinformatics, University of Leeds Press, UK.

Langevin, P. (1905), Magnetisme et Theorie des Electrons, *Annual of Chemistry Physics*, **5**, 71-127.

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

Mardia, K. V. (1975), Statistics of Directional Data (with discussion), *Journal of Royal Statistical Society*, B, **37**, 349-393.

Mardia, K. V. and Jupp, P. E. (2000), *Directional Statistics*, John Wiley, Chichester.

Mardia, K. V., Taylor, C. C. and Subramaniam, G. K. (2007), Protein Bioinformatics and Mixture of Bivariate Von Mises Distributions for Angular Data, *Biometrics*, **63**, 505-512.

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

Mardia, K. V. (2009), Statistical Complexity in Protein Bioinformatics, In Mardia, K. V., Gusnanto, A. and Fallaize, C. J. Editors, *Statistical Tools for Challenges in Bioinformatics*, University of Leeds Press, UK.

Mardia, K. V. (2010), Bayesian Analysis for Bivariate Von Mises Distributions, *Journal of Applied Statistics*, **37**, 515-528.

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

Rivest, L. P. (1988), A Distribution for Dependent Unit Vectors, *Communications in Statistics*, **17**, 461-483.

Schmidt, J. W. and R. E. Taylor, (1970), *Simulation and Analysis of Industrial Systems*, Richard D. Irwin, Homewood, Illinois.

Singh, H., Hnizdo, V. and Demchuk, E. (2002), Probabilistic Model for Two Dependent Circular Variables, *Biometrika*, **89**, 719-723.

Von Mises, R. (1918), Uber die "Ganzzahligkeit" der Atomgewichte und Verwandte Fragen, *Physikal*, **19**, 490-500.