مجله علوم آماری، پاییز و زمستان ۱۳۹۳
جلد ۸، شماره ۲، ص ۲۲۳-۲۴۳

نحوه خوشه‌بندی آماری داده‌های شکل

مهناز نبیل، موسی گل‌علی‌زاده
گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۲/۱۱/۲۳ تاریخ آخرین بازنگری: ۱۳۹۳/۶/۲۶

چکیده : اخیرا به‌کارگیری ابزارهای آمار چندمتغیره برای تحلیل داده‌هایی که به‌صورت هندسی تصادفی هستند مورد اقبال محققین علوم کاربردی قرار گرفته است. آمارشکل به عنوان شاخه جدیدی از هندسه تصادفی شامل مجموعه‌ای از چنین داده‌هایی است. با این حال، چون چنین داده‌هایی ماهیت غیراقلیدسی دارند نحوه تطبیق ابزارهای مرسوم چندمتغیره برای تحلیل آماری مناسب آن‌ها تا حدودی واضح نیست. در این مقاله نحوه خوشه‌بندی داده‌های آمارشکل مطالعه شده، سپس عملکرد آن با رویکرد مرسوم آمار چندمتغیره به این موضوع در قالب تحلیل مثال کاربردی مرتبط با استخوان فمور ران مورد مقایسه قرار می‌گیرد.

واژه‌های کلیدی : خوشه‌بندی، هندسه تصادفی، تحلیل آماری شکل، فواصل آماری، استخوان فمور ران.

۱ مقدمه

داده‌های چندمتغیره که شامل متغیرهای کمی، کیفی یا تلفیقی از آن‌ها باشند را می‌توان توسط مدل‌های متنوعی شامل تحلیل چندمتغیره پیوسته و گسسته مورد

آدرس الکترونیک مسئول مقاله: موسی گل‌علی‌زاده، golalizadeh@modares.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲P۱۰، ۳۲C۰۵

مطالعه قرار داد. شاخه جدیدی از آمار که برای تحلیل داده‌های چندمتغیره مدنظر گرفته شده، مطالعه آماری ساختار هندسی داده‌ها است که به آمارشکل[۱] معروف شد (کندال، ۱۹۸۴؛ اسمال، ۱۹۹۶؛ درایدن و ماردیا، ۱۹۹۸). یکی از شروط اصلی تحلیل در این حوزه، حفظ ساختار هندسی اولیه در تمامی مراحل تحلیل آماری است. ابزارهای معرفی شده در این شاخه تحت نام تحلیل آماری شکل با فعالیت‌های ارزشمند کندال و بوک استاین در دهه ۷۰ وارد حوزه آمار شده است (کندال، ۱۹۷۷؛ بوک استاین، ۱۹۷۸). تعریف رسمی (بر اساس مفهوم ریاضی) شکل به‌عنوان سنگ‌بنای تحلیل آماری شکل، توسط کندال (۱۹۷۷) به‌صورت زیر ارائه شد:

> شکل تمامی اطلاعات هندسی حاصل از حذف اثرات دوران، مکان و مقیاس است که از یک شیء باقی می‌ماند.

در روش‌های استاندارد آمارشکل، ابتدا تعدادی نقاط شاخص روی کران اطراف شیء قرار داده شده و مختصات آنها توسط دستگاه مختصات دلخواهی (عموماً دکارتی یا قطبی) تعیین می‌شوند. سپس مراحل از بین بردن اثرات انتقال، مقیاس و دوران روی مجموعه نقاط انتخابی اجرا و بر اساس نوع تبدیلات اعمال شده، که منجر به قرارگیری داده‌ها در فضایی موسوم به فضای شکل می‌گردد، تحلیل‌های آماری متناسب با آنها صورت می‌گیرند. لازم به اشاره است که در این مقاله اخذ نمونه تنها از قسمت‌های بیرونی اشیاء مورد مطالعه قرار گرفته و اخذ اطلاعات از بخش درونی آنها، که از موضوعات نوین تحقیق آمارشکل است (دورمن و همکاران، ۲۰۱۳)، صرف‌نظر شده است. نکته مهم و اساسی مرتبط با تحلیل آمارشکل این است که داده‌های تبدیل شده در فضایی غیرخطی (نااقلیدسی) قرار می‌گیرند (پنک، ۱۹۹۹). به‌همین دلیل، داشتن متری مناسب برای اندازه‌گیری فاصله بین شکل‌ها و دسته‌بندی مجموعه‌ای از شکل‌ها، از ضروریات اساسی تحلیل داده‌های شکل است.

خوشه‌بندی داده‌های شکل با استفاده از ابزارهای آمار چندمتغیره از موضوعات مورد توجه در حوزه آمار کاربردی است. اما، منابع موجود در این حوزه تنها محدود

---

[۱] Shape statistics

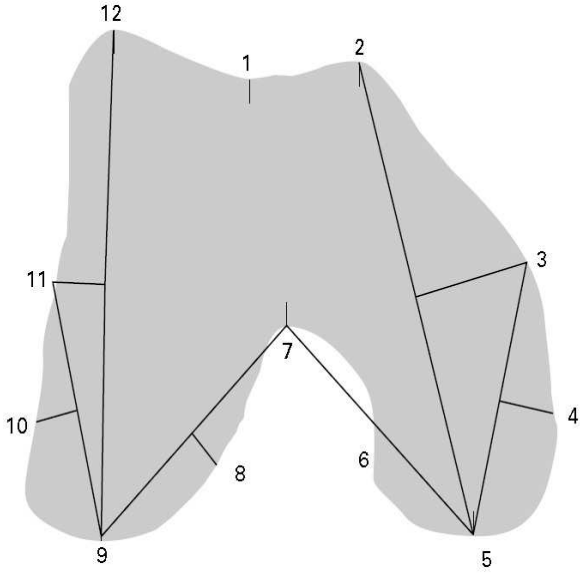
به سریواستاوا و همکاران (۲۰۰۵)، آمارال و همکاران (۲۰۱۰) و ایشیهارا و همکاران (۲۰۱۱) است. با این حال، به‌جز منبع دوم دو منبع دیگر از ابزارهای آماری غیر شکل برای خوشه‌بندی اشیاء استفاده کردند. به‌طور کلی، آنچه که می‌توان از دید محقق آمار چندمتغیره برای خوشه‌بندی داده‌های شکل عنوان کرد، این است که کلیدی‌ترین ابزار برای این مهم، داشتن معیاری جامع از فاصله است. شاید فواصل آماری شکل برای معرفی واریانس شکل مناسب باشند اما برای دسته‌بندی آن‌ها بر اساس روش‌های موجود خوشه‌بندی، از کارایی لازم برخوردار نیستند. این موضوع ناشی از ماهیت غیراقلیدسی داده‌های شکل است. لذا، هدف مقاله حاضر مطالعه نحوه خوشه‌بندی آماری داده‌های شکل و مقایسه آن با رویکرد مرسوم آمار چندمتغیره است.

در این مقاله خلاصه‌ای از مفاهیم آمارشکل و بعضی از فواصل مناسب داده‌های شکل در بخش ۲ ارائه می‌شود. سپس، نحوه انجام آزمون برابری میانگین‌های شکل در بخش ۳ تشریح می‌شود تا توسط آن ضرورت خوشه‌بندی داده‌های شکل مورد بررسی قرار گیرد. آنگاه، در بخش ۴ خوشه‌بندی داده‌های مورد مطالعه با دو رویکرد عدم حفظ ساختار هندسی (تحلیل استاندارد چندمتغیره) و حفظ ویژگی هندسی اشیاء (تحلیل در فضای شکل) صورت گرفته و در نهایت تفاوت نتایج حاصل از این دو رویکرد مورد ارزیابی قرار می‌گیرد. نتیجه‌گیری کلی نیز در انتهای مقاله آمده است. بیان این نکته ضروری است که کلیه محاسبات آماری این مقاله با استفاده از بسته shapes در نرم افزار آماری R نسخه ۳.۱.۰ انجام شده است.

۲ مقدمه‌ای بر آمارشکل

شپستون و همکاران (۱۹۹۹) در تحقیقی تصویر مقطع بالایی از ۶۸ استخوان فمور ران پا را مورد بررسی قرار دادند. آن‌ها با قرار دادن دوربینی در فاصله‌ای مشخص از مقطع بالایی تمامی فمورها، تصویرهای مورد نیاز را به‌دست آورده و سپس با قرار دادن ۱۲ نقطه روی قسمت بیرونی سطح مقطع، مختصات دکارتی دو بعدی نقاط انتخابی را ثبت کردند. تصویر یکی از استخوان‌ها و مکان هندسی نقاط شاخص

ثبت شده در شکل ۱ آمده است.



شکل ۱: تصویر مقطع بالایی استخوان فمور ران و مکان نقاط شاخص روی آن

داده‌ها دارای دو متغیر رسته‌ای نیز هست که یکی از آنها، متغیر "سمت" است که طرف پای متناظر با فمور را نشان می‌دهد و مقدار آن می‌تواند معرف پای چپ یا راست باشد. دیگری متغیر "وضعیت" است که وضعیت سطح مقطع استخوان از جهت مقدار غضروف باقی‌مانده را نشان می‌دهد که به دو گروه عاجی‌شده[۲] و عاجی‌نشده[۳] تقسیم می‌شود.

برای تحلیل داده‌های شکل، ابتدا استانداردسازی آن‌ها بر روی مختصات دکارتی‌شان صورت می‌گیرد. برای این منظور، توابع ریاضی متفاوتی وجود دارند که به‌کارگیری هر کدام منجر به مختصات مختلفی از شکل می‌شود. مرسوم‌ترین مختصات از بین آن‌ها عبارتند از مختصات کندال[۴] (کندال، ۱۹۸۴)، بوک استاین[۵]

۲ Eburnated
۳ Non-Eburnated
۴ Kendall coordinates
۵ Bookstein coordinates

(بوک استاین، ۱۹۸۶)، و پروکراستس[۶] (گودال، ۱۹۹۱). جزئیات محاسباتی کلی برای محاسبه این مختصات در درایدن و ماردیا (۱۹۹۸) موجود است.

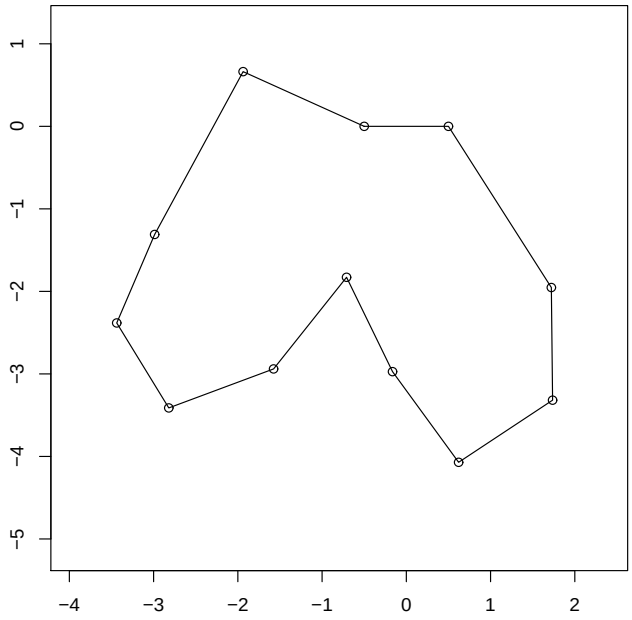
مختصات بوک استاین از ساده‌ترین مختصات شکل بوده و برای شخصی که بخواهد به لحاظ چشمی با آمارشکل ارتباط برقرار کند، بسیار مفید است، چرا که تعداد دیگری از مختصات مثل کندال نمایش هندسی واضحی ندارند. به زبانی دقیق‌تر، برای نمایش هندسی اشکال با استفاده از آن مختصات باید از تصویر کره سه بعدی یا با بعدهای بالاتر در صفحه مختصات دکارتی استفاده کرد که در برخی از مواقع امری آسان نیست. کندال و همکاران (۱۹۹۹) سعی کردند با استفاده از هندسه تصادفی و مفاهیم توپولوژی نحوه نمایش در کره را ارائه کنند. البته از نقطه نظر ریاضی وقتی که اشیاء در فضای دو بعدی قرار دارند تناظر یک به یکی بین مختصات مطرح شده وجود دارد که در آن صورت مجددا نقش مختصات بوک استاین به دلیل نمایش مستقیم اشکال در صفحه دکارتی برجسته‌تر خواهد شد. اما باید اشاره شود که برای مقصود مقاله حاضر، مختصات پروکراستس از اهمیت بیشتری نسبت به بقیه برخوردار است. شاید بتوان گفت یکی از این دلایل مربوط به وجود معیاری واقعی از فاصله بین شکل‌ها در تعریف چنین مختصات از داده‌های شکل است. با این حال، برای درک بهتری از داده‌های مربوط به ۶۸ استخوان فمور ران نمایش هندسی مختصات بوک استاین میانگین شکل (M_{Book}) آن‌ها در شکل ۲ آمده است.

بنا به ماردیا و همکاران (۱۹۷۹) تحلیل پروکراستس دربرگیرنده مراحل انطباق ماتریسی مثل Y بر روی ماتریس دیگری مثل X است تا جایی که کمیت

$$\sum_{r=1}^{n}(x_r - cA'y_r - b)'(x_r - cA'y_r - b)$$

که به فاصله پروکراستس معروف است، کمینه شود به قسمی که x_r و y_r به‌ترتیب سطرهای ماتریس‌های X و Y هستند. حال اگر ماتریس‌های X و Y حاوی مختصات دو شیء باشند، آنگاه براساس این تحلیل به شیء اول اجازه داده می‌شود به سمت شیء دوم تغییر مکان پیدا کرده، در صورت نیاز اندازه خود را با آن یکی

[۶] Procrustes coordinates



شکل ۲: نمودار میانگین بوک استاین شکل فمور

نموده و تا جایی دوران پیدا کند که بتواند بر آن منطبق شود. مختصات حاصل از اعمال این تبدیلات روی شیء مورد مطالعه مختصات پروکراستس نامیده می‌شود. توجه شود که فاصله پروکراستس برای حالتی که بیش از دو شیء در دسترس باشند نیز قابل تعمیم است. در آمارشکل فاصله پروکراستس در حالت اول را معمولی[۷] و در حالت دوم را تام[۸] می‌گویند.

از میان فواصل آماری شکل که در منابع یاد شده آمده، فاصله پروکراستس تام در فضای مماسی[۹] محاسبه می‌شود. فضای مماسی، صفحه مماس بر ابرکره داده‌های شکل است. استفاده از داده‌های تصویر شده در این فضا امکان استفاده از روش‌های معمول آمار چندمتغیره را به محقق می‌دهد. اما شرط استفاده از داده‌های شکل در فضای مماسی، تمرکز مطلوب آنها در فضای شکل است. این فرض در بسیاری از

۷ Ordinary

۸ Full

۹ Tangent space

موارد معقول نیست و استفاده از آن منجر به بهم‌ریختگی ساختار هندسی اشیاء می‌شود. در این موارد معیارهای دیگری از فاصله در فضای شکل مطرح می‌شوند که یکی از آنها فاصله ریمانی[۱۰] است. این فاصله، طول کمان دایره گذرنده از دو شکل روی ابرکره داده‌های شکل است. جزئیات بیشتری از این فواصل و ارتباط آن‌ها در درایدن و ماردیا (۱۹۹۸) آمده است.

واضح است که برای دستیابی به میانگین‌های متفاوت شکل داشتن مبنایی برای ارزیابی تفاوت و اختلاف بین داده‌های شکل ضروری است. برای مختصات بوک استاین متوسط فاصله اقلیدسی بین مختصات اشیاء و میانگین بوک استاین، مبنایی برای فاصله است. در حالی‌که براساس مختصات پروکراستس، این معیار توسط فاصله پروکراستس تام به‌دست می‌آید. علاوه بر این‌ها، فواصل دیگری نیز در آمارشکل وجود دارند که جزئیات بیشتری از آن‌ها در اسمال (۱۹۹۶) و درایدن و ماردیا (۱۹۹۸) وجود دارند. به‌علاوه، بررسی موضوع محاسبه میانگین در فضاهای نااقلیدسی شامل فضای شکل، در کندال و همکاران (۱۹۹۹) و لی و کومه (۲۰۰۰) آمده است.

برای خوشه‌بندی داده‌های شکل باید بتوان تفکیک دقیقی از میانگین شکل گروه‌ها به عمل آورد. بخش بعد آزمون برابری میانگین‌های شکل را به قصد خوشه‌بندی داده‌های شکل ارائه می‌کند.

۳ آزمون برابری میانگین‌های شکل

قبل از تشریح مراحل آزمون برابری میانگین‌های شکل جزئیات چنین عملی برای داده‌های چندمتغیره، که قصدی برای حفظ ساختار هندسی آن‌ها در میان نباشد، ارائه می‌شود. جزئیاتی از این آزمون‌ها در اکثر کتب استاندارد چندمتغیره مانند ماردیا و همکاران (۱۹۷۹) موجود است.

فرض کنید X_1 و X_2 ماتریس داده‌های مستقل $n_i \times p$ بعدی و به‌ازای $i = 1, 2$ تمامی n_i سطر ماتریس X_i مستقل از هم و هر یک

[۱۰] Reimannian

دارای توزیع $N_p(\mu_i, \Sigma_i)$ باشند. آنگاه اگر $\mu_۲ = \mu_۱$ و $\Sigma_۲ = \Sigma_۱$، آماره $Q = (n_۱ n_۲/(n_۱ + n_۲ - ۲))(\bar{X}_۱ - \bar{X}_۲)' S_u^{-۱} (\bar{X}_۱ - \bar{X}_۲)$ دارای توزیع هتلینگ $T^۲(p, n_۱ + n_۲ - ۲)$ است طوری‌که $\bar{X}_i$ میانگین نمونه iام و

$$S_u = (n_۱ S_۱ + n_۲ S_۲)/(n_۱ + n_۲ - ۲)$$

که در آن S_i ماتریس کوواریانس نمونه iام است. بنا به آمار چندمتغیره، چون $T^۲(p, m) = \frac{mp}{m-p+۱} F_{p, m-p+۱}$، می‌توان بر اساس آماره Q و توزیع F، برابری میانگین دو جامعه مستقل را آزمود.

در صورت نابرابری ماتریس‌های کوواریانس یا نرمال نبودن توزیع داده‌ها، انجام آزمون به‌صورت ناپارامتری اجتناب‌ناپذیر است. فرض کنید $X_{۱i}$ و $X_{۲i}$ برای $i = ۱, \ldots, n_j$ و به‌ازای $j = ۱, ۲$ نمونه‌های تصادفی مستقل از جامعه‌هایی با تابع توزیع تجمعی $F(x)$ و $G(x)$ باشند. $\bar{X}$ را بردار میانگین نمونه حاصل از ترکیب این دو مجموعه و زوایایی که بردارهای $X_{۱i} - \bar{X}$ و $X_{۲i} - \bar{X}$ با جهت مثبت محور xهامی‌سازند به‌صورت بردارهای رتبه $(r_۱, \ldots, r_{n_۱})$ و $(r'_۱, \ldots, r'_{n_۲})$ در نظر بگیرید. برای آزمون فرضیه $H_۰ : F(x) = G(x)$ در مقابل $H_۱ : F(x) = G(x + \delta)$ به‌ازای $\delta \neq ۰$ از آماره

$$U = \frac{۲(N-۱)}{n_۱ n_۲} \{ (\sum_{i=۱}^{n_۱} \cos \frac{۲\pi r_i}{N})^۲ + (\sum_{i=۱}^{n_۲} \sin \frac{۲\pi r_i}{N})^۲ \}$$

استفاده می‌شود، که در آن $N = n_۱ + n_۲$. وقتی $n_۱, n_۲ \to \infty$ و $\frac{n_۱}{n_۲} \to \beta$ که $۰ < \beta < ۱$، آنگاه تحت $H_۰$ به‌طور مجانبی داریم: $U \sim \chi^۲_۲$. در نتیجه $H_۰$ به‌ازای مقادیر بزرگ U رد می‌شود.

برای به‌کارگیری این آزمون‌ها راجع به داده‌های مقاله حاضر، مقادیر مختصات نقاط شاخص به‌عنوان متغیر در نظر گرفته شده‌اند. بنابراین ماتریس داده‌ها، ماتریسی با بعد ۲۴ × ۶۸ است که بنا به متغیر سمت به دو گروه راست و چپ، هر یک شامل ۳۴ مشاهده تقسیم شده، ولی بنا به متغیر وضعیت به دو گروه با تعداد ۵۲ و ۱۶ تایی به‌ترتیب برای حالت‌های عاجی‌شده و عاجی‌نشده تقسیم می‌شوند. نتیجه این آزمون به دو صورت پارامتری و ناپارامتری در جدول ۱ آمده است طوری‌که

برای اجرای آزمون ناپارامتری، تعداد بازنمونه‌گیری ۱۰۰۰ اختیار شده است.

جدول ۱: p–مقدار آزمون مقایسه چندمتغیره میانگین‌ها

	روش آزمون	
متغیر	پارامتری	ناپارامتری
سمت	$< e^{-16}$	$< e^{-16}$
وضعیت	۰/۰۱۸۷۷	۰/۰۱۷۴

بنا به این جدول و براساس هر دو نوع آزمون، میانگین‌های دو گروه چپ و راست متفاوت ولی میانگین‌های دو گروه عاجی‌شده و عاجی‌نشده یکسان هستند. بنابراین انتظار می‌رود تحلیل خوشه‌ای داده‌های استخوان فمور ران با ابزار آمار چندمتغیره و بدون لحاظ ویژگی هندسی آن‌ها تنها قادر به دسته‌بندی داده‌ها به دو گروه چپ و راست بوده ولی نمی‌تواند آن‌ها را بنا به ویژگی عاجی بودن دسته‌بندی کند.

اکنون مقایسه میانگین گروه‌ها براساس هر یک از متغیرها و البته با مدنظر قرار دادن ساختار هندسی به‌عنوان یک ویژگی مؤثر مدنظر قرار می‌گیرد. آزمون‌های متفاوتی برای ارزیابی برابری میانگین‌های شکل وجود دارند (داریدن و ماردیا، ۱۹۹۸). ساده‌ترین و در عین حال مرسوم‌ترین روش، تصویر مختصات از فضای شکل به فضای مماسی و سپس اجرای آزمون دو نمونه‌ای مستقل T^2 هتلینگ است (کنت، ۱۹۹۴). واضح است که شرط مورد نیاز برای اجرای این آزمون، نرمال بودن مختصات تصویر شده و هم‌واریانسی بین دو گروه مورد بررسی است.

آزمون گودال[11] یکی دیگر از آزمون‌های آمارشکل برای مقایسه میانگین شکل‌های دو نمونه مستقل است (گودال، ۱۹۹۱). فرض مورد نیاز برای اجرای این آزمون، کوچک بودن واریانس مشترک مختصات‌هاست. این فرض در منابع آمارشکل به فرض همسانگردی ماتریس‌های کوواریانس دو گروه معروف است. ممکن است در بعضی از موارد واریانس دو جامعه مورد مقایسه با هم فرق داشته

[11] Goodall

باشند. در این صورت آماره جیمز[۱۲] ابزار مناسبی برای آزمون تساوی دو میانگین شکل است (آمارال و همکاران، ۲۰۰۷). در بعضی موارد نادر توصیه شده است که بدون محدود شدن به فضای غیراقلیدسی شکل، مقایسه میانگین‌ها را براساس نسبت درستنمایی تعمیم یافته (کسلا و برگر، ۲۰۰۲) انجام داد. آماره حاصل از این رویکرد به آماره لامبدا معروف است (آمارال و همکاران، ۲۰۰۷). در ادامه خلاصه‌ای از این آزمون‌ها می‌آید.

۱.۳ آزمون T^2 هتلینگ

دو نمونه تصادفی $X_1,\ldots,X_{n_1}$ و $Y_1,\ldots,Y_{n_2}$ از دو جامعه مستقل با میانگین‌های شکل $[\mu_1]$ و $[\mu_2]$ را در نظر بگیرید، طوری که هر کدام از X_i و Y_j به ازای $i=1,\ldots,n_1,\ j=1,\ldots,n_2$ نمایانگر ماتریس‌های پیکره‌بندی شامل k نقطه شاخص در m بعد هستند. برای آزمون $H_0 : [\mu_1]=[\mu_2]$ در مقابل $H_1 : [\mu_1]\neq[\mu_2]$ ابتدا مشاهدات شکل به فضای مماسی تصویر می‌شوند. توجه کنید که در این حالت، $\hat{\mu}$ میانگین پروکراستس مشاهدات تلفیق شده دو نمونه است.

اگر مقادیر مختصات پروکراستس متناظر، به‌ترتیب با $v_1,\ldots,v_{n_1}$ و $w_1,\ldots,w_{n_2}$ نشان داده شود، به‌دلیل قرار داشتن در فضای اقلیدسی می‌توان توزیع آنها را نرمال در نظر گرفت. به این ترتیب، فرض می‌شود:

$$v_i\sim N(\xi_1,\Sigma),\ i=1,\ldots,n_1,\quad w_j\sim N(\xi_2,\Sigma),\ j=1,\ldots,n_2.$$

مشابه آزمون T^2 هتلینگ در حالت چندمتغیره استاندارد، در این حالت از آماره $H^2=(\bar{v}-\bar{w})'S_u^-(\bar{v}-\bar{w})$ استفاده می‌شود که $\bar{v}$ و $\bar{w}$ میانگین‌های دو نمونه و S_u^- معکوس مون-پنروز واریانس نمونه‌ای ادغام شده است. تحت فرض صفر، آماره

$$F \;=\; \frac{n_1n_2(n_1+n_2-M-1)}{(n_1+n_2)(n_1+n_2-2)M}H^2 \qquad (1)$$

دارای توزیع F_{M,n_1+n_2-M-1} است طوری‌که $M=km-m-1-m(m-1)/2$. توجه شود که M نمایانگر بعد داده‌های شکل است.

[۱۲] James

در صورتی که فرض برابری ماتریس‌های کوواریانس یا لحاظ توزیع نرمال برای مختصات پروکراستس برقرار نباشد، از معادل ناپارامتری آزمون (۱) استفاده می‌شود. در این حالت داده‌ها در دو گروه با حجم یکسان جایگشت شده و آماره آزمون برای تمامی B جایگشت ممکن محاسبه می‌گردد. آن‌گاه، p–مقدار آزمون $\frac{r-1}{B} - 1$ خواهد بود که r رتبه آماره آزمون براساس داده‌های مشاهده شده است (کنت، ۱۹۹۴).

۲.۳ آزمون F گودال

برای آزمون میانگین‌های شکل، می‌توان از آماره‌هایی براساس توان دوم فواصل پروکراستس و توزیع آماری مرتبط با این فواصل استفاده کرد (گودال، ۱۹۹۱). برای تشریح دقیق‌ترآزمون F گودال، فرض کنید مشاهدات شکل از جوامعی با مدل خطی

$$X_i = \beta_i(\mu + E_i)\Gamma_i + 1_k\gamma_i', \quad vec(E_i) \sim N(0, \sigma^2 I_{km})$$

تولید شده‌اند که $\beta_i > 0$، $\Gamma_i \in SO(m)$ و $\gamma_i \in R^m$ به‌ترتیب پارامترهای مقیاس، دوران و مکان بوده و vec عملگر برداری‌کننده است. در اینجا X_i ماتریسی $k \times m$ بعدی حاوی مختصات شیء است. γ_i ماتریس $m \times 1$، E_i و μ ماتریس‌های $k \times m$ بعدی هستند.

تحت فرض صفر $[\mu_1] = [\mu_2]$ و با قبول فرض برابری ماتریس کوواریانس‌های برای دو مجموعه ماتریس‌های پیکره‌بندی، گودال (۱۹۹۱) ثابت کرده است

$$\begin{aligned}
\sum_{i=1}^{n_1} d_F^2(X_i, \hat{\mu}_1) &\sim \tau_0^2 \chi^2_{(n_1-1)M}, \\
\sum_{i=1}^{n_2} d_F^2(Y_i, \hat{\mu}_1) &\sim \tau_0^2 \chi^2_{(n_2-1)M}, \\
d_F^2(\hat{\mu}_1, \hat{\mu}_2) &\sim \tau_0^2 (\frac{1}{n_1} + \frac{1}{n_2}) \chi^2_M.
\end{aligned}$$

که در آن کمیت‌های سمت چپ دو به دو از هم مستقل هستند. به‌علاوه، $\tau_0 = \sigma/\delta_0$ و $\delta_0 = S(\mu_0)$ که δ_0 اندازه مرکزی‌شده μ_0 است که با استفاده از رابطه

$S(X) = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X}_j)^2}$ به‌دست می‌آید به‌قسمی که $\bar{X}_j = \frac{1}{k}X_{ij}$. در نهایت، آماره F گودال به‌صورت

$$F = \frac{n_1 + n_2 - 2}{n_1^{-1} + n_2^{-1}} \frac{d_F^2(\hat{\mu_1}, \hat{\mu_2})}{\sum_{i=1}^{n_1} d_F^2(X_i, \hat{\mu}_1) + \sum_{i=1}^{n_2} d_F^2(Y_i, \hat{\mu}_1)} \qquad (2)$$

است، که از توزیع $F_{M,(n_1+n_2-2)M}$ پیروی می‌کند و مقادیر بزرگ آن منجر به عدم تائید فرض صفر می‌شود. توجه شود که M همان مقدار در رابطه (۱) است.

در صورت عدم برقرای مفروضات آزمون، از حالت ناپارامتری آن استفاده می‌شود. انجام آزمون ناپارامتری در این حالت مشابه جزئیات تشریح شده در انتهای آزمون هتلینگ است، با این تفاوت که در این حالت باید آماره مدنظر براساس رابطه (۲) محاسبه شود.

۳.۳ آزمون جیمز

اگر فرض برابری ماتریس‌های کوواریانس برای دو مجموعه ماتریس‌های پیکره‌بندی برقرار نباشد و شخص در پی آزمونی پارامتری برای مقایسه میانگین‌های شکل باشد، می‌تواند از گونه اصلاح شده آزمون هتلینگ که به آزمون جیمز معروف است (آمارال و همکاران، ۲۰۰۷)، استفاده نماید. در این حالت فرض نرمال بودن مشاهدات کماکان از ضروریات آزمون است.

با همان نمادگذاری مرسوم دو زیر بخش قبلی آماره آزمون جیمز عبارتست از:

$$F_J = (\bar{v} - \bar{w})'(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2)^{-}(\bar{v} - \bar{w}). \qquad (3)$$

همان‌طور که انتظار می‌رود در این حالت، $F_J \sim \chi_M^2$. در صورت عدم برقرای مفروضات این آزمون، از حالت ناپارامتری آن استفاده می‌شود، که انجام آن نیز مشابه جزئیات تشریح شده در انتهای آزمون هتلینگ است، با این تفاوت که در این حالت آماره آزمون بر اساس رابطه (۳) محاسبه می‌شود.

اکنون می‌توان آزمون برابری میانگین‌های شکل را برای داده‌های استخوان فمور ران بکار گرفت. نتیجه آزمون برابری میانگین‌ها شامل p–مقدارهای آزمون با

استفاده از جداول آماری مرتبط و همچنین بر اساس بازنمونه‌گیری در جدول ۲ ارائه شده است. همان‌طور که ملاحظه می‌شود نمی‌توان در سطح ۰/۰۱ تفاوتی بین میانگین‌های شکل دو گروه پای چپ و راست قائل شد. لذا، مشابه آنچه که در برداشت عامیانه در خصوص تقارن بین پاهای افراد وجود دارد، ملاحظه می‌شود که شکل استخوان‌های چپ و راست (از نقطه نظر هندسی) افراد مورد مطالعه با هم یکسان است.

جدول ۲: p–مقدار آزمون مقایسه میانگین‌های شکل

آزمون	سمت	وضعیت
هتلینگ	۰/۴۵۷۵	۰/۳۷۳۱
	۰/۴۲۵۷	۰/۲۹۷۰
جیمز	۰/۳۳۷	۰/۳۹۸
	۰/۴۰۵۹	۰/۷۵۲۵
گودال	۰/۹۹۷۵	۰/۹۶۱۲
	۰/۱۹۹۰	۰/۱۵۴۲

به‌طور مشابه می‌توان آزمون مقایسه میانگین‌های شکل را برای داده‌های استخوان براساس متغیر "وضعیت" انجام داد. به‌عبارتی دیگر، هدف مقایسه متوسط شکل دو گروه عاجی‌شده و نشده است. نتایج محاسباتی برای این فرضیه در ستون آخر جدول ۲ آمده است. همان‌طور که ملاحظه می‌شود p–مقدارها از فرضیه برابری میانگین شکل دو گروه حمایت می‌کنند.

نتیجه کلی این بخش نشان می‌دهد که تفاوتی بین میانگین‌های شکل استخوان‌های فمور ران چپ و راست و همچنین استخوان‌های عاجی‌شده و نشده وجود ندارد. لذا، انتظار می‌رود تحلیل خوشه‌ای استخوان‌ها با کمک آمارشکل قادر به دسته‌بندی استخوان پاهای چپ و راست و همچینن عاجی‌شده و نشده نباشد. به‌عبارتی دیگر، شواهد ماحصل از آمار چندمتغیره مبنی بر تفاوت بین دو گروه، بدون توجه به ساختار هندسی استخوان‌هاست وگرنه آن‌ها از نظر شکلی شبیه هم هستند.

۴ تحلیل خوشه‌ای داده‌های استخوان

در این قسمت برای تحلیل خوشه‌بندی داده‌های استخوان از روش سلسله مراتبی استفاده شده است. همچنین، دو رویکرد آمارشکل و آمار چندمتغیره مورد مقایسه قرار گرفت تا بتوان درک بهتری از تحلیل آمارشکل ارائه نمود. علاوه بر این، سعی شد هنگام تحلیل چندمتغیره استاندارد با اتخاذ دو رویکرد متفاوت در منظور نمودن داده‌ها در ماتریس مشاهدات، میزان حساسیت خوشه‌بندی را به‌طریقی غیرمستقیم مورد ارزیابی قرار داد. لذا، مجموعه مختصات نقاط به دو شیوه ذیل وارد تحلیل شدند:

- ورود مختصات ۱۲ نقطه شاخص بر اساس مختصات دکارتی x و y
- ورود مختصات x به‌طور جداگانه بدون در نظر گرفتن مختصات y و برعکس

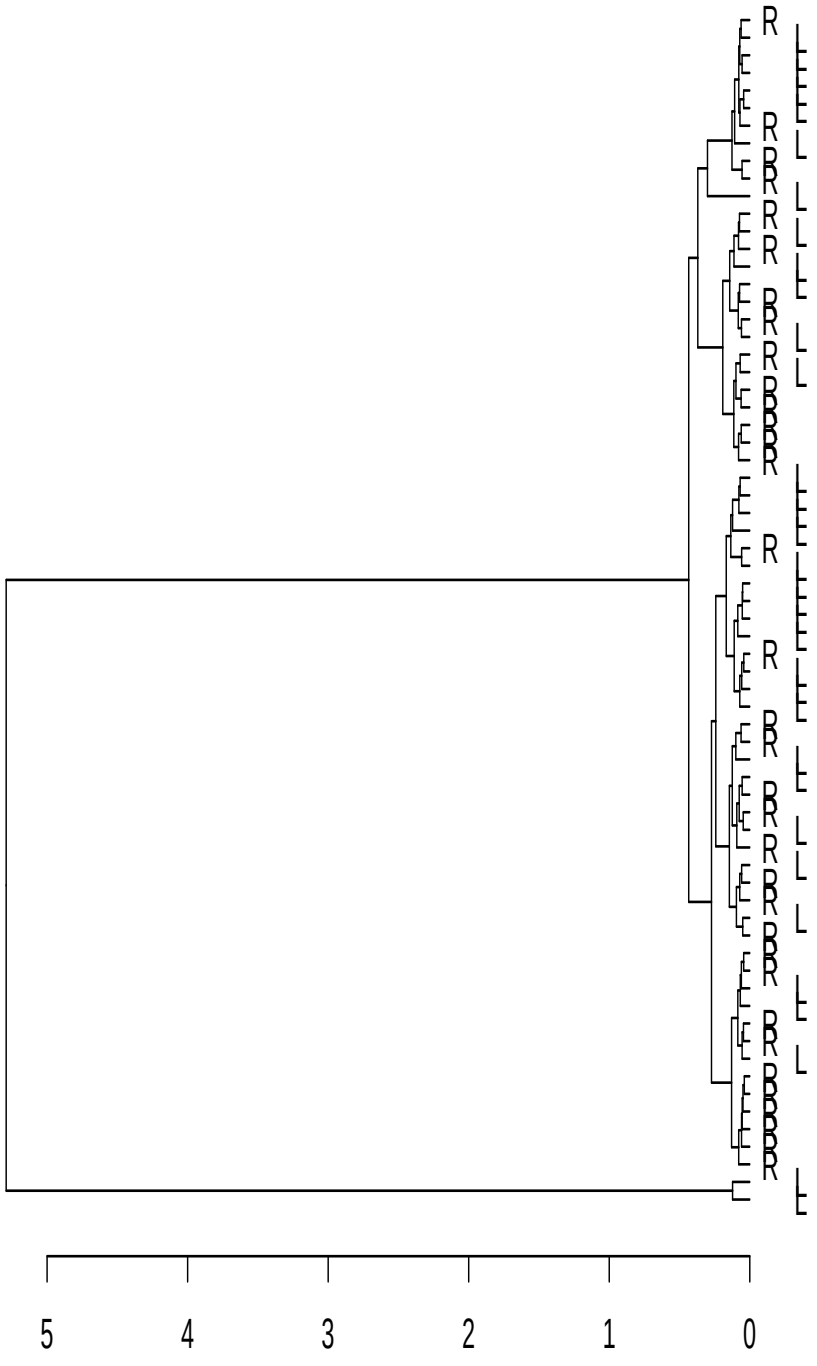
به‌علاوه تمامی مختصات به‌صورت مؤلفه به مؤلفه استاندارد شدند. در نهایت تلاش شد فواصل متفاوت موجود در آمار چندمتغیره مورد استفاده قرار گیرد تا دید وسیع‌تری از لحاظ خوشه‌بندی داده‌های مورد مطالعه، توسط آمار چندمتغیره به‌دست آید.

یک نمایش شماتیک از اجرای خوشه‌بندی برای درک بهتر از تحلیل‌های صورت گرفته، مفید خواهد بود. به‌خصوص خوشه‌بندی داده‌های شکل، دریچه جدیدی از تحلیل چندمتغیره با رویکردی نوین ارائه می‌کند. نتیجه خروجی یکی از روش‌ها در شکل ۳ آمده است. بیان این نکته ضروری است که برای یکپارچه‌سازی خروجی حاصل از نرم‌افزار R تعدیل بخشی از درختواره ضروری می‌نمود. مثلا، کلمات R و L که به‌ترتیب معرف استخوان فمور پای راست و چپ هستند، بزرگتر از اندازه خودشان رسم شده‌اند تا وضوح درختواره بیشتر شود. با این‌حال، در بعضی از موارد، نمایش حاصل مطلوب نشده است. علی‌رغم این موضوع، همان‌طور که از شکل ۳ ملاحظه می‌شود این برچسب‌ها به‌صورت نامنظم پراکنده شده‌اند. به‌عبارتی دیگر، نمی‌توان تفکیک دقیقی بین مشاهدات موجود بر حسب دو ویژگی متفاوت متغیر "سمت" قائل شد. علاوه بر این، اندازه خوشه‌ها نیز ناهمگون هستند در

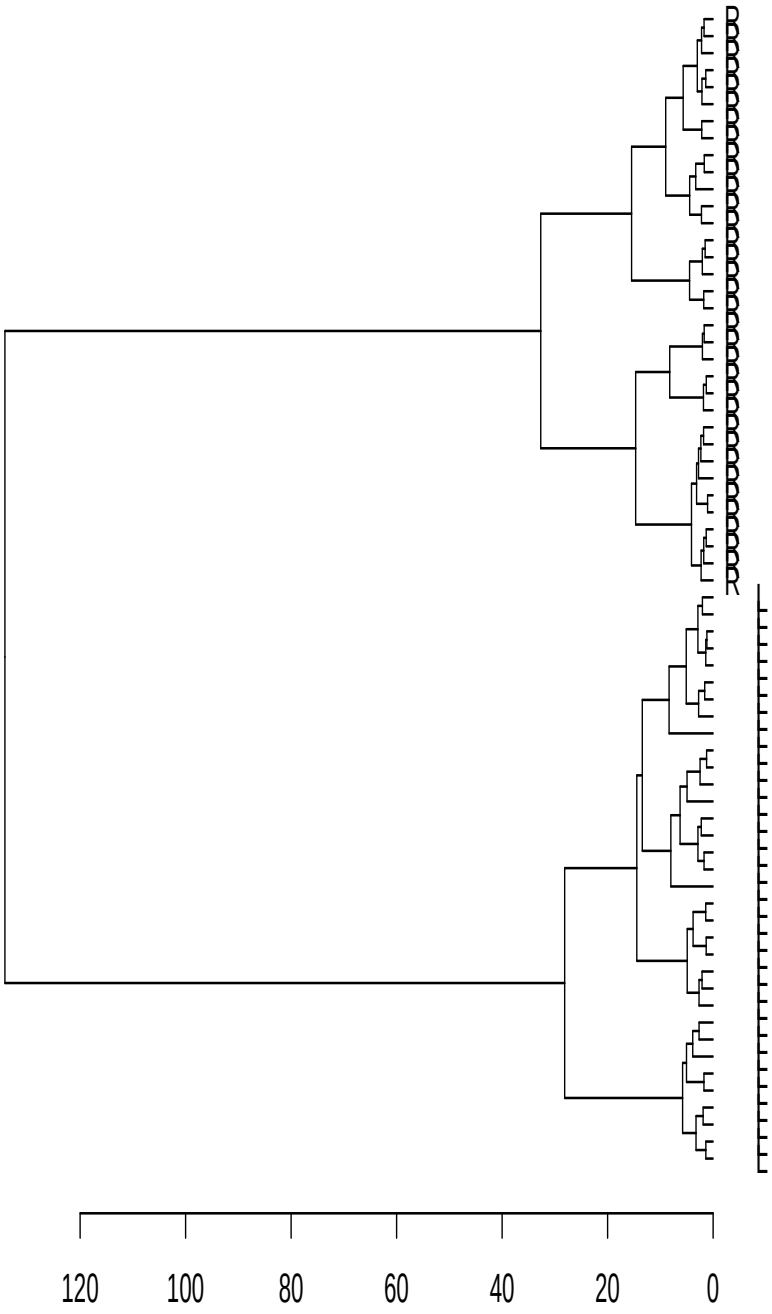
حالی‌که تعداد برچسب‌ها برابرند. تفاوت ناهمگون خوشه‌ها که به طریقی منعکس کننده فاصله نمونه‌ها از هم هستند، نیز قابل تأمل است.

برای مقایسه چشمی تفاوت استفاده از رویکرد آمار چندمتغیره و آمارشکل در خوشه‌بندی داده‌های استخوان، یکی از نمودار درختواره حاصل از اجرای روش سلسله مراتبی با استفاده از متر اقلیدسی اما بدون لحاظ ویژگی هندسی داده‌ها در شکل ۴ آمده است. ملاحظه می‌شود خوشه‌ها در دو دسته مجزا تفکیک شده‌اند. این امر با توجه به معنی‌دار شدن تفاوت بین میانگین دو گروه در حالت چندمتغیره که نتایج آن در جدول ۱ ارائه شده بود، قابل انتظار است. به عبارتی دیگر، وقتی که میانگین‌های با برچسب راست و چپ با هم فرق دارند، آنگاه طبیعی است که آن‌ها می‌بایست در دو دسته مجزا قرار گیرند.

به‌طور کلی تقابل بین آمارشکل و آمار چندمتغیره استاندارد از این حیث حائز اهمیت است که معرفی و ارزیابی معیارهای نوینی از فاصله برای هر کدام از این دو رویکرد ضروری است. به عبارت دیگر، به‌دست آوردن معیارهایی از یک رویکرد برای رویکرد دیگری و مقایسه نتیجه حاصل با یک محدودیت منطقی نخواهد بود. همان‌طور که مشاهده شد چون شکل دو گروه داده براساس متغیرهای "سمت" و "وضعیت" یکسان بودند، استفاده از فاصله‌های بین اشکال که در منابع آمارشکل مثل داریدن و ماردیا (۱۹۹۸) وجود دارد، داده‌های شکل مورد مطالعه را به حق خوشه‌بندی می‌کند. از طرف دیگر، یکی از دلایل خوشه‌بندی مناسب داده‌های استخوان توسط رویکرد آمار چندمتغیره، وجود تفاوت طبیعی بین مختصات نقاط شاخص در دو گروه است نه شکل آن‌ها. لذا، شکل ۴ تفاوت بین مختصات در هر کدام از نمونه داده‌ها را به نمایش گذاشته است. بنابراین می‌توان نتیجه گرفت که اگر شخصی بخواهد داده‌های مورد مطالعه‌اش را با درنظر گرفتن ویژگی هندسی، و به‌طور دقیق‌تر بر اساس شکل، آنها دسته‌بندی کند، باید از اطلاعات شکل، با مفهومی که کندال (۱۹۷۷) ارائه کرده است، برای این منظور استفاده نماید. استفاده صرف از مقادیر مختصات اشیاء منجر به نتایج مناسبی نخواهد شد. به زبانی ساده، ممکن است دو شیء با مختصات متفاوت، دربردارنده اطلاعات هندسی یکسانی باشند. به‌عنوان مثال، اگرچه مختصات شکل صورت یک فرد که توسط نقاش و



شکل ۳: نمودار درختواره خوشه‌بندی داده‌های استخوان با آمارشکل



شکل ۴: نمودار درختواره خوشه‌بندی داده‌های استخوان با آمار چندمتغیره

عکاس ثبت می‌شود با هم فرق دارند اما شکل آن‌ها یکسان بوده و همان فرد مورد نظر را معرفی می‌کنند.

نکته مورد علاقه محققین در اجرای تحلیل خوشه‌بندی، مقایسه روش‌های متفاوت با استفاده از برخی معیارهای مناسب کمی است. با اغماض از تقابل مورد اشاره در بالا (در مورد آمارشکل و آمار استاندارد چندمتغیره)، یک معیار کمی می‌تواند محدودیتی را برای انتخاب بهترین روش خوشه‌بندی از بین روش‌های موجود، ارائه کند. تعدادی از این معیارها در هندل و همکاران (۲۰۰۵) آمده است. یکی از آن‌ها معیار $Dunn$ است که برای یک روش خوشه‌بندی مثل C، از رابطه

$$D(C) = \frac{\min_{C_k, C_l \in C}\{\min_{i \in C_k, j \in C_l} dist(i,j)\}}{\max_{C_m \in C} diam(C_m)}$$

به‌دست می‌آید، که در آن C_i خوشه iام، $diam$ ماکسیمم فاصله بین مشاهدات در خوشه و $dist$ فاصله مورد استفاده است. مقادیر معیار $D(C)$ از صفر تا بی‌نهایت تغییر می‌کند و هرچه این مقدار بیشتر باشد، روش خوشه‌بندی موردنظر مطلوب‌تر است. مقدار این معیار برای تمامی روش‌های ادغامی خوشه‌بندی منجر به عددی بسیار مشابه شد. لذا، نمی‌توان یکی از روش‌ها را نسبت به دیگری ترجیح داد.

بحث و نتیجه‌گیری

در مقاله حاضر جزئیاتی از نحوه خوشه‌بندی داده‌هایی که معرف شکل هستند ارائه گردید و از برخی فواصل مرسوم آمارشکل برای اجرای تحلیل خوشه‌بندی کمک گرفته شد. نتایج نشان داد که ماهیت آمارشکل و به‌ویژه فضای نااقلیدسی آن دقت بیش از پیشی را طلب می‌کند. به‌عنوان مثال، اگرچه ممکن است تفکیک گروه‌ها توسط آزمون فرضیه‌های آماری تأیید گردند، لیکن نمی‌توان انتظار چنین تفکیکی را از تحلیل خوشه‌بندی داشت. لذا، بنا به ویژگی توپولوژیکی فضای شکل، تحلیل خوشه‌بندی برای آمارشکل نیازمند مطالعه و معرفی تعمیم ویژه‌ای از آن است.

تحقیقات آتی در این حوزه از آمارشکل می‌تواند به‌صورت‌های متفاوتی بسط داده شود. ممکن است به‌کارگیری گونه دیگری از فواصل آماری برای داده‌های شکل بتواند توصیف مناسب‌تری از نحوه خوشه‌بندی ارائه کند. یکی از این فواصل،

فاصله اندازه–شکل[۱۳] است. مطالعه تحلیل خوشه‌بندی با این رویکرد می‌تواند مدنظر علاقه‌مندان به این حوزه از آمار قرار گیرد. استفاده از معیارهای فاصله‌ی نااقلیدسی دیگر مثل فاصله ژئودزیک (فلتچر و همکاران، ۲۰۰۳؛ فلتچر و همکاران، ۲۰۰۴) و امکان‌سنجی روش‌های دیگر تحلیل خوشه‌بندی مثل روش مدل‌مبنا نیز برای تحقیقات آتی حائز توجه هستند. علاوه بر این، بررسی تمامی این موضوعات با رویکرد آمار بیزی نیز می‌تواند مورد توجه محققین علاقه‌مند به این حوزه قرار گیرد.

تقدیر و تشکر

نویسندگان مقاله از نظرات سازنده و مفید داوران محترم که در بهبود مقاله حاضر بسیار موثر بوده است، کمال تشکر را دارند.

مراجع

Amaral, G. J .A., Dore, L. H., Lessa, R. P. and Stosic, B. (2010), K-means Algorithm in Statistical Shape Analysis, *Communications in Statistics - Simulation and Computation*, **39**, 1016-1026.

Amaral, G. J .A., Dryden, I. L. and Wood, A. T. W. (2007), Pivotal Bootstrap Methods for K-sample Problems in Directional Statistics and Shape Analysis, *Journal of the American Statistical Association*, **102**, 695-707.

Bookstein, F. L. (1978), *The Measurment of Biological Shape and Shape Change*, Lecture Notes on Biomathematics, Springer, New York.

[۱۳] Size-and-shape distance

Bookstein, F. L. (1986), Size and Shape Spaces for Landmark Data in Two Dimensions (with discussion), *Statistical Sciences*, **1**, 181-242.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, 2nd Ed., Thomson Learning, Duxbury.

Dryden, I. L. and Mardia, K. V. (1998), *Statistical Shape Analysis*, John Wiley, Chichester.

Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G. and Ayache, N. (2013), Toward a Comprehensive Framework for the Spatio Temporal Statistical Analysis of Longitudinal Shape Data, *International Journal of Computer Vision*, **103**, 22-59.

Fletcher, P. T., Joshi, S., Lu, C. and Pizer, S. M. (2004), Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape, *IEEE Transactions on Medical Imaging*, **23**, 995-1005.

Fletcher, P. T., Lu, C. and Joshi, S. (2003), Statistics of Shape via Principal Geodesic Analysis on Lie Groups, In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 95-101.

Goodall, C. R. (1991), Procrustes Methods in the Statistical Analysis of Shape (with discussion), *Journal of the Royal Statistical Society*, B, **53**, 285-339.

Handel, J., Knowles, J. and Kell, D. B. (2005), Computational Cluster Validation in Post-Genomic Data Analysis, *Bioinformatics*, **21**, 3201-3212.

Ishihara, S., Ishihara, K. and Nagamachi, M. (2011), Statistical Shape Analysis of Headlights, In *Proceedings of IEEE International Confer-*

ence on Biometrics and Kansei Engineering, 27-32.

Kendall, D. G. (1977), The Diffusion of Shape, *Advances in Applied Probability*, **9**, 428-430.

Kendall, D. G. (1984), Shape Manifolds, Procrustean Metrics and Complex Projective Spaces, *Bulletin of the London Mathematical Society*, **16**, 81-121.

Kendall, D. G., Barden, D., Carne, T. K. and Le, H. L. (1999), *Shape and Shape Theory*, John Wiley, Chichester.

Kent, J. T. (1994), The Complex Bingham Distribution and Shape Analysis, *Journal of the Royal Statistical Society*, B, **56**, 285-299.

Le, H. L. and Kume, A. (2000), The Fréchet Mean and the Shape of the Means, *Advances in Applied Probability*, **32**, 101-114.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*, Academic Press, London.

Pennec, X. (1999), Probabilities and Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements, In *Proceedings of Nonlinear Signal and Image Processing*, 194-198.

Shepstone, L., Rogers, J., Kirwan, J. R. and Silverman, B. (1999), The Shape of the Distal Femur: A Comparison Between Eburnated and Non-Eburnated, *Annals of the Rheumatic Diseases*, **58**, 72-78.

Small, G. (1996), *The Statistical Theory of Shape*, Springer, New York.

Srivastava, A., Joshi, S. H., Mio, W. and Liu, X. (2005), Statistical Shape Analysis: Clustering, Learning and Testing, *IEEE Transactions On Pattern Analysis and Machine Intelligence*, **27**, 590-602.