

بهبود عملکرد تحلیل ژنودزیک اصلی در تحلیل آمار شکل

حمیدرضا فتوحی، موسی گل علی زاده

گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۱/۳/۱۸ تاریخ آخرین بازنگری: ۱۳۹۱/۹/۶

چکیده: یکی از هدف‌های تحلیل آماری شکل، علاوه بر دستیابی به برآوردی از میانگین، برآورد واریانس شکل است، که از طریق تحلیل مولفه اصلی قابل حصول است. به دلیل محدودیت استفاده از روش تحلیل مولفه اصلی برای مجموعه داده‌هایی از فضای اقلیدسی، این روش برای داده‌های آمار شکل که ماهیتاً غیراقلیدسی هستند، قابل کاربرد نیست. در این حالت می‌توان از تحلیل ژنودزیک اصلی یا تقریب خطی آن به عنوان تعمیمی از تحلیل مولفه اصلی به فضای غیراقلیدسی استفاده نمود. از آنجا که اساس این روش مبتنی بر الگوریتم گرادیان نزول است، در این مقاله با آشکار ساختن چند ضعف عمده آن، الگوریتم جدیدی معرفی می‌شود که هم منجر به برآورد استوار میانگین شکل و هم باعث حفظ ساختار هندسی شکل خواهد شد. سپس با ارائه جنبه‌های نظری روش تحلیل ژنودزیک اصلی، عملکرد آن در یک مطالعه شبیه‌سازی و یک مثال واقعی مورد ارزیابی قرار خواهد گرفت.

واژه‌های کلیدی: تحلیل مولفه اصلی، تحلیل ژنودزیک اصلی، الگوریتم گرادیان نزول، فضای غیراقلیدسی شکل.

پیشرفت‌های اخیر تکنولوژی موجب تسهیل مطالعه اشیاء از روی تصاویر هندسی آنها شده است. این موضوع علم آمار را با مسائل متفاوت و جدیدی مواجه کرده است که شکل شیء یک منبع اطلاعاتی مناسب به شمار می‌رود.

کندال (۱۹۸۴) و بوک استاین (۱۹۸۶) از جمله کسانی بودند که به‌طور همزمان به مطالعه آماری اشیاء با در نظر گرفتن ساختار هندسی آنها پرداختند. حاصل زحمات آنها و محققین بعد از آنها، منجر به پیدایش شاخه جدیدی به نام آمار شکل گردیده است. یکی از جنبه‌های مهم تحلیل آماری شکل، برآورد ساختار تغییرات مربوط به شکل، از طریق نمونه تصادفی اخذ شده از اشیاء است. برای محاسبه واریانس شکل کوئز و همکاران (۱۹۹۲) و کنت (۱۹۹۴) برای اولین بار استفاده از تحلیل مولفه اصلی را در فضای مماس، که فضایی اقلیدسی است، به عنوان تقریبی از فضای شکل پیشنهاد کردند. این تقریب به این دلیل است که فضای شکل غیراقلیدسی است و استفاده مستقیم از تحلیل مولفه‌های اصلی در آن امکان‌پذیر نخواهد بود (درایدن و ماردیا، ۱۹۹۸).

پیروی از این روش ممکن است در بعضی موارد خاص قابل توجیه باشد، اما بهتر است واریانس شکل بدون تقریب و مستقیماً در همان فضای شکل محاسبه شود. لذا به نظر می‌رسد تعمیمی از تحلیل مولفه اصلی به فضای غیراقلیدسی، راه‌حل این مسئله باشد. یکی از تعمیم‌های تحلیل مولفه اصلی به فضای غیراقلیدسی تحلیل ژئودزیک اصلی^۱ (PGA) است که اولین بار توسط فلتچر (۲۰۰۴) برای تحلیل تصاویر کامپیوتری پیشنهاد شد. اخیراً سامر و همکاران (۲۰۱۰) به ارزیابی این روش در چند مثال خاص پرداخته‌اند. مقاله حاضر با نگرشی اجمالی به این تحقیقات، الگوریتم جدیدی برای محاسبه میانگین استوار شکل با حفظ ساختار هندسی در مراحل نیل به میانگین ارائه می‌کند.

در بخش دوم این مقاله مطالبی مختصر از آمار غیراقلیدسی که دربرگیرنده آمارشکل نیز است ارائه می‌شود. در بخش سوم دو معیار برای ارزیابی تغییرات

^۱ Principal Geodesic Analysis

داده‌های آمار شکل تشریح و عملکرد آن‌ها در یک مثال کاربردی مورد مطالعه قرار می‌گیرد. در بخش چهارم به منظور بهبود عملکرد معیارهای ذکر شده، الگوریتم جدیدی پیشنهاد و کارایی آن در یک مطالعه شبیه‌سازی و یک مثال واقعی با حالت استاندارد مورد مقایسه قرار می‌گیرد. سپس نتیجه‌گیری و پیشنهاداتی برای تحقیقات آتی در بخش پنجم ارائه خواهد شد.

۲ آشنایی با آمار غیراقلیدسی

در سال‌های اخیر مطالعه داده‌هایی که فضای رخداد آن‌ها غیراقلیدسی هستند مورد توجه محققین قرار گرفته است (سامر و همکاران، ۲۰۱۰). دایره و کره مثال‌هایی از این قبیل هستند که ماردیا و جاپ (۲۰۰۰) از آنها به عنوان داده‌های سوئی یاد می‌کنند. بر اساس مفاهیم توپولوژیکی، در منابع علمی چنین فضاها را منیفلد^۲ (M) می‌نامند. به زبانی ساده، منیفلد تعمیمی از روبه در فضاها سه‌بعدی و بیشتر است. یکی از مهم‌ترین موضوعات آماری مربوط به مطالعه داده‌ها در منیفلدی خاص، نحوه محاسبه خلاصه‌های آماری است.

مشابه محاسبه میانگین داده‌ها در فضای اقلیدسی، پس از تعیین متر ریمانی مربوط به فضای منیفلد، که معمولاً آن را با d نمایش می‌دهند، با محاسبه مسئله کمینه‌سازی به صورت

$$\mu = \arg \min_{x \in M} \sum_{i=1}^N d^2(x, x_i) \quad (1)$$

و تعیین جواب آن در صورت وجود، میانگینی تحت عنوان میانگین ذاتی نتیجه می‌شود (پنس، ۱۹۹۹).

الگوریتم گرادیان نزول^۳ (GDA)، یکی از پرکاربردترین روش‌های محاسبات عددی برای نیل به این هدف است (کوشی، ۱۸۴۷؛ لئونبرگر، ۱۹۶۹؛ مزا، ۲۰۱۰). این الگوریتم در برگیرنده دو پارامتر تاثیرگذار، به نام‌های خطای الگوریتم (ε) و اندازه گام برداری (τ) است و معمولاً توسط کاربر تعیین می‌شود.

^۲ Manifold

^۳ Gradient Descent Algorithm

محاسبه واریانس در فضاهای غیراقلیدسی ارتباط تنگاتنگی با محاسبه میانگین در آن فضا دارد. به زبانی دقیق‌تر، واریانس نمونه‌ایی مشاهدات متعلق به یک فضای غیراقلیدسی را می‌توان به عنوان مجموع توان‌های دوم فاصله ریمانی مشاهدات از میانگین ذاتی به صورت

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N d^2(\mu, x_i) = \frac{1}{N} \sum_{i=1}^N \|\log_{\mu}(x_i)\|^2,$$

تعریف کرد، که در آن $\|\cdot\|$ نرم اقلیدسی، همچنین $\log_{\mu}(x_i)$ بیانگر نگاهی به منظور انتقال مشاهده x_i از فضای غیراقلیدسی به فضای مماس با مرکزیت μ بوده و به نگاشت لگاریتمی معروف است (فلتچر و همکاران، ۲۰۰۴). توجه کنید که نماد $\exp_{\mu}(x_i)$ معکوس نگاشت لگاریتمی است و از آن به عنوان نگاشت نمایی یاد می‌شود.

ویژگی اصلی داده‌های آمار شکل این است که غیراقلیدسی هستند، یعنی داده‌های حاصل از شکل ساختار هندسی اشکال دویعدی، در یک ابر کره قرار دارند. در حالت خاص، اگر شیء مورد مطالعه مثلث باشد، کندال (۱۹۸۴) نشان داد فضایی که داده‌های برآمده از شکل در آن قرار دارند کره‌ای با شعاع نیم است.

۳ ارزیابی تغییرات داده‌های آمار شکل

پس از محاسبه میانگین، برآورد واریانس یکی از مسائل مورد علاقه در تحلیل آماری شکل است. برای دستیابی به این هدف، کنت (۱۹۹۴) استفاده از تحلیل مولفه اصلی در فضای مماس را پیشنهاد نمود.

فرض کنید n ماتریس $(k+1) \times m$ بعدی X_1, \dots, X_n به عنوان پیکره‌بندی n شیء مختلف در اختیار است. با پیش ضرب این ماتریس‌ها توسط زیر ماتریس هلمرت H با بعد $(k+1) \times k$ و سپس تقسیم آنها به نرم اقلیدسی اشیاء مورد نظر به ماتریس‌های جدیدی می‌رسیم که نسبت به تبدیلات انتقال و مقیاس پایا هستند. آنها را به عنوان اطلاعات پیش شکل با نمایش ماتریسی Z_1, \dots, Z_n در نظر بگیرید. در این صورت مختصات متغیرهای شکل در فضای مماس با مرکزیت میانگین

شکل را می‌توان توسط رابطه

$$v_j = (I_{k-m} - \text{Vec}(\hat{\mu})\text{Vec}(\hat{\mu})^T)(\text{Vec}(Z_i\hat{\Gamma}_i)) \quad j = 3, \dots, k$$

تقریب زد، که $\text{Vec}(\cdot)$ بیانگر عملگر برداری کننده، $\hat{\mu}$ برآورد میانگین شکل به روش پروکراستس تام و $\hat{\Gamma}$ برآورد ماتریس دوران براساس انطباق بهینه پیش شکل‌ها به روش پروکراستس تام است (کنت، ۱۹۹۴).

اکنون ماتریس کواریانسِ مختصات شکل در فضای مماس را به صورت $S_v = \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^T$ در نظر بگیرید. واضح است که به کمک S_v اعمال روش تحلیل مولفه اصلی در فضای مماس امکان‌پذیر خواهد بود. به‌ویژه، اثر زامین مولفه بر روی مشاهده i ام به صورت

$$x_{ij} = \gamma_j^T (v_i - \bar{v}) \quad i = 1, \dots, n; \quad j = 1, \dots, p \leq n$$

به‌دست می‌آید، که در آن γ_j بردارهای ویژه ماتریس کواریانس S_v و λ_j مقدارهای ویژه متناظر با γ_j است. همچنین می‌توان با تقسیم زامین مولفه بر مقدار ویژه متناظرش، به اثر استاندارد شده j امین مولفه بر مشاهده i ام از طریق رابطه

$$\sigma_{ij} = x_{ij} / \lambda_j^{1/2}, \quad i = 1, \dots, n; \quad j = 1, \dots, p$$

دست یافت. حال با توجه به اینکه هر مولفه اصلی بیانگر درصدی از تغییرات کل می‌باشد، درصدی از عمده تغییرات شکل، پیرامون میانگین جامعه، تحت هر مولفه قابل محاسبه است. به‌ویژه با محاسبه میانگین شکل، معمولاً تغییرات اصلی شکل تحت هر مولفه از طریق کمیت $v = \bar{v} + \sigma_{ij} \lambda_j^{1/2} \gamma_j$ به‌ازای $\sigma_{ij} = \pm 3$ توصیف می‌شود.

هرچند می‌توان روش تحلیل مولفه اصلی را برای محاسبه واریانس در فضای شکل، که غیراقلیدسی است، مورد استفاده قرار داد اما یکی از مشکلات اساسی محدود بودن این روش به فضای اقلیدسی است. لذا استفاده مستقیم آن در فضای شکل امکان‌پذیر نبوده و همان‌گونه که کنت (۱۹۸۴) پیشنهاد نمود، باید فضای شکل را با فضای مماس، که اقلیدسی است، تقریب زد، سپس روش تحلیل مولفه اصلی را

به کار بست. پیروی از این روش ممکن است در بعضی موارد خاص قابل توجیه باشد، اما بهتر است واریانس شکل بدون تقریب و مستقیماً در فضای غیر اقلیدسی شکل محاسبه شود.

PGA مشابه تحلیل مولفه اصلی، علاوه بر کاهش بعد داده‌ها روشی برای برآورد واریانس جامعه است. باید توجه داشت که برخلاف تحلیل مولفه اصلی که استفاده مستقیم آن محدودیت‌هایی را در بر دارد، می‌توان آن را زمانی که فضای مورد مطالعه غیر اقلیدسی است به کار برد. به بیانی دیگر این روش تعمیمی از تحلیل مولفه اصلی به فضای غیر اقلیدسی است. لذا هنگام بکارگیری آن، با فرض غیر اقلیدسی بودن فضای مورد مطالعه، هدف تصویر داده‌ها به داخل زیر فضایی با بعد کمتر است، طوری که داده‌های تصویر شده دارای بیشترین پراکندگی در اطراف میانگین باشند. محاسبه جهت‌های ژنودزیک اصلی در فضای شکل نیازمند محاسبه فرم کلی نگاهت لگاریتمی برای ابر کره با نقطه دلخواه μ به عنوان نقطه شروع است. فتوحی و گل علی‌زاده (۲۰۱۲) نشان دادند که با تعمیم معادله صفحه مماس روی کره‌ایی با ابعاد بالا ($S^d, d > 3$)، می‌توان شکل کلی نگاهت لگاریتمی نقطه $\vec{x} = (x_1, \dots, x_d) \in S^d$ با شروع از $\vec{\mu}_0 = (x_{01}, \dots, x_{0d})$ را به صورت

$$\log_{\vec{\mu}}(\vec{x}) = \frac{\theta}{\sin(\theta)}(\vec{x}_{[-d]}, 1) - \lambda \vec{\mu}$$

نوشت، که در آن $\theta = \arccos(x_d)$ و

$$\lambda = \frac{\theta}{\sin(\theta)}(\vec{\mu}_{[-d]} \times \vec{x}_{[-d]}) + x_{0d} - 1$$

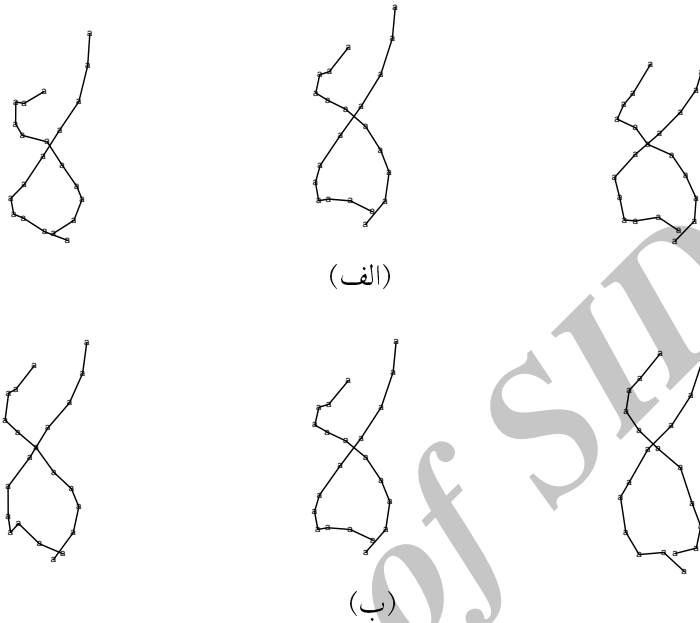
است، به قسمی که $\vec{\mu}_{[-d]} = (x_{01}, \dots, x_{0d-1})$ و $\vec{x}_{[-d]} = (x_1, \dots, x_{d-1})$ بیان این نکته ضروری است که این مقاله شامل دو سری مجموعه داده‌ها است. داده‌های اول مربوط به حرکت یک مولکول DNA در ۳۰ زمان متفاوت می‌باشد، که اطلاعات مربوط به هر زمان در قالب ماتریسی با بعد 22×3 ذخیره شده است. داده‌های مثال دوم، اطلاعات حاصل از تصاویر پزشکی مغز تعدادی از افراد است (درآیدن و همکاران، ۲۰۰۲؛ فری و همکاران، ۲۰۰۱). این داده‌ها از طریق پایگاه الکترونیکی <http://cran.r-project.org/web/packages/shape> قابل دسترس است.

در این بخش برای بررسی عملکرد تحلیل مولفه اصلی و PGA، مجموعه داده‌های اول مدنظر قرار گرفته می‌شود. برای برآورد ساختار تغییرات مربوط به حرکت مولکول DNA، ابتدا فضای شکل با فضای مماس تقریب زده شده و دو مولفه اصلی اول در شکل ۱ رسم شده است که در آن ردیف اول بیانگر عمده پراکندگی پیرامون میانگین شکل تحت اولین مولفه اصلی است. در واقع از روی این ردیف و با توجه به مقدار ویژه متناظر با اولین جهت اصلی، می‌توان نتیجه گرفت که در حدود ۴۴/۲ درصد از پراکندگی کل مربوط به نقاط شاخصی است که قسمت میانی بین دو شاخک مولکول و پیرامون آن را تشکیل می‌دهند. همچنین با توجه به ردیف دوم می‌توان دریافت که ۱۲/۶۸ درصد از تغییرات کل مربوط به قسمت‌های انتهایی شاخک‌های تشکیل دهنده مولکول است.

به منظور توصیف دقیق‌تر واریانس حرکت DNA در فضای واقعی شکل و بدون وجود محدودیت حاصل از به‌کارگیری تحلیل مولفه اصلی، از PGA برای ارزیابی تغییرات مربوط به حرکت مولکول بهره‌جسته و دو ژنودزیک تقریبی اول در شکل ۲ رسم شده است، ردیف اول بیانگر عمده پراکندگی پیرامون میانگین شکل تحت اولین ژنودزیک اصلی است. در حقیقت از روی این ردیف و با توجه به مقدار ویژه متناظر با اولین جهت ژنودزیک اصلی، می‌توان نتیجه گرفت که در حدود ۷۹/۹ درصد از پراکندگی کل مربوط به میزان کشیدگی مولکول است. همچنین ردیف دوم این مطلب را بیان می‌کند که ۳/۸۴ درصد از تغییرات کل مربوط به نقاط شاخص میانی و نقاط انتهایی است. واضح است که تحقیق بیشتر در مورد حرکت مولکول DNA نیازمند لحاظ کردن نظر متخصصان زیست‌شناسی و در عین حال مدل‌بندی جامع‌تر است.

۴ ارزیابی عملکرد الگوریتم گرادیان نزول

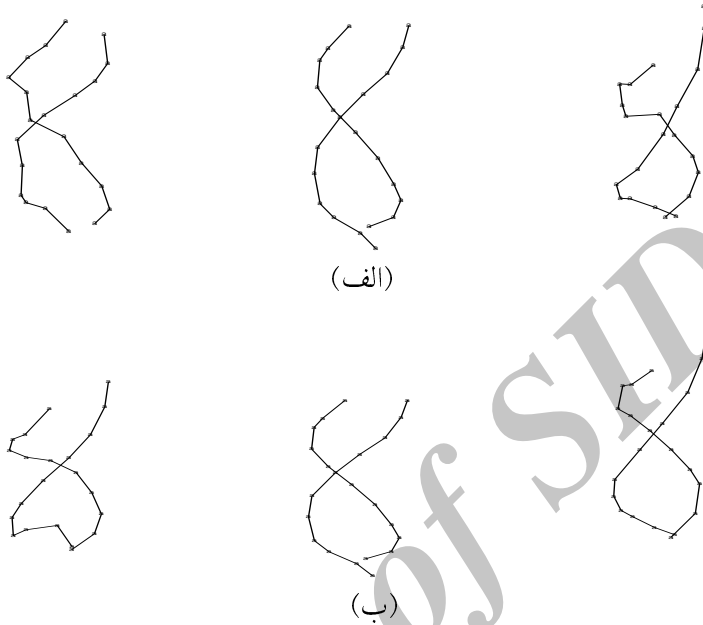
با مطالعه چگونگی دستیابی به میانگین ذاتی شکل می‌توان دریافت که GDA نقش اساسی را در فرآیند تکراری بازی می‌کند. با توجه به ماهیت این الگوریتم، هنگام استفاده از تحلیل ژنودزیک اصلی، ممکن است زمان رسیدن به میانگین ذاتی شکل



شکل ۱: واریانس ساختار مولکول DNA حول میانگین (نمودار وسط) با ± 3 انحراف استاندارد (نمودار چپ و راست). تغییرات توصیف شده توسط اولین مولفه (نمودار الف) ۴۴/۲ درصد و دومین مولفه (نمودار ب) ۱۲/۶۸ درصد

طولانی شود یا اساساً منجر به نتیجه مطلوبی نشود. برای روشن شدن این مطلب و بیان جزئیات بیشتر در این خصوص، ابتدا به بررسی تاثیر τ در عملکرد الگوریتم مطرح شده پرداخته و سپس با استفاده از داده‌های واقعی آمارشکل، نقش آن در ایجاد ضعف برای GDA نشان داده خواهد شد. بنا به محاسبات شبیه‌سازی و تجربیات نویسندگان این مقاله، باید به این نکته اشاره کرد که انتخاب مقدار مناسب ε تا حدودی سلیقه‌ای است و آنچه تاثیر بیشتری در عملکرد الگوریتم دارد پارامتر τ است. از این رو بیشتر به نقش τ معطوف خواهد شد، اما در بخش مثال واقعی با انتخاب چند مقدار مختلف برای ε نقش این پارامتر در دستیابی به میانگین ذاتی شکل نیز نشان داده خواهد شد.

در این قسمت عملکرد GDA در یک مثال کاربردی، مربوط به تصاویر پزشکی مغز تعدادی از افراد مورد بررسی قرار می‌گیرد و نشان داده خواهد شد GDA از



شکل ۲: واریانس ساختار مولکول DNA حول میانگین (نمودار وسط) با ± 3 انحراف استاندارد (نمودار چپ و راست). تغییرات توصیف شده توسط اولین ژنودزیک (نمودار الف) $79/9$ درصد و دومین ژنودزیک (نمودار ب) $3/84$ درصد

چند ضعف عمده برخوردار است. سپس الگوریتم پیشنهادی جدیدی برای بهبود آن ارائه می‌شود.

بنا به توضیحات مربوط به GDA، ضروری است مقادیر ε و τ از قبل مشخص شوند. از آنجا که هدف بررسی تاثیر این دو پارامتر بر عملکرد آن است، ε برابر $0/1, 0/4, 0/7, 1, \tau, 0/01, 0/06, 0/1, 0/6$ اختیار شده است. به علاوه میانگین حسابی همه ماتریس‌های پیکره‌بندی موجود به عنوان نقطه شروع اولیه برای الگوریتم در نظر گرفته شده است. برای مقادیر ε و τ میانگین ذاتی شکل محاسبه و تعداد تکرارهای ممکن تا رسیدن به همگرایی در قالب جدول ۱ ارائه شده‌اند. البته در بعضی از موارد که الگوریتم واگرا است به جای تعداد تکرار، کلمه «واگرا» یادداشت شده است.

بنا به جدول ۱، به جز برای $\varepsilon = 0/6$ ، وقتی مقدار τ به ۱ می‌رسد GDA واگرا می‌شود. این موضوع برای $\varepsilon = 0/01$ حتی برای مقادیر کم τ نیز اتفاق می‌افتد. نکته دیگر همگرایی GDA در تکرارهای دوم برای مقادیر τ غیر از $0/1$ و $0/4$ است. پیش بینی می‌شود برای ε های بزرگ‌تر، تعداد تکرار تا رسیدن به همگرایی عددی بیشتر از ۲ باشد. نتیجه کلی این جدول، به نوعی بیانگر عدم ثبات رفتار توسط الگوریتم برای مقادیر متفاوت ε و τ است. به‌ویژه در مواقعی که این روش همگرا نمی‌شود، نقش این دو پارامتر برجسته‌تر از وضعیت‌هایی است که روش به همگرایی می‌رسد.

جدول ۱: تعداد تکرار در اجرای GDA تا رسیدن به همگرایی.

| | ε | | | | τ |
|--|---------------|-------|-------|-----|--------|
| | 0/01 | 0/06 | 0/1 | 0/6 | |
| | ۱ | ۱ | ۱ | ۱ | 0/1 |
| | واگرا | ۲ | ۱ | ۱ | 0/4 |
| | واگرا | ۲ | ۲ | ۲ | 0/7 |
| | واگرا | واگرا | واگرا | ۲ | ۱ |

برای بررسی وضعیت ساختار هندسی شکل حاصل از گام‌های مختلف این روش، میانگین پروکراستس که به آسانی قابل محاسبه است به عنوان ملاک ارزیابی، مورد استفاده قرار گرفته است. از این‌رو از ملاک OPA^۴ برای مقایسه دو ساختار هندسی استفاده شده است (فتوحی و گل‌علی‌زاده، ۲۰۱۲).

بر اساس این ملاک دو ساختار هندسی طوری روی هم‌دیگر منطبق می‌شوند که کمترین اختلاف ممکن را از یک‌دیگر داشته باشند. جزئیات بیشتر در زمینه انطباق بهینه در درآیدن و ماردیا (۱۹۹۸) موجود است. برای مثال ذکر شده، ساختار هندسی اول، میانگین پروکراستس شکل و ساختار هندسی دوم، ماتریس پیکره‌بندی خروجی حاصل از هر مرحله الگوریتم در نظر گرفته شد. مقدار τ همان مقادیر فوق و ε برابر $0/06$ در نظر گرفته شد و برای هر حالت کمیت OPA محاسبه شده

^۴ Ordinary Procrustes Analysis

است. نتایج حاصل، همراه با نمایش هندسی دو میانگین پروکراستس و ذاتی در شکل ۳ نمایش داده شده است. همان‌گونه که ملاحظه می‌شود با افزایش τ کمیت OPA نیز افزایش می‌یابد. در نتیجه اختلاف بین دو میانگین بیشتر می‌شود، و تعبیر آن این است که ساختار هندسی میانگین ذاتی شکل خیلی دورتر از میانگین پروکراستس شکل خواهد شد. به عبارتی دیگر ناپایداری در ساختار هندسی بیشتر شده و میانگین حاصل از الگوریتم ذکر شده معتبر نخواهد بود.

نتایج کلی این بخش به ما کمک می‌کند تا الگوریتم جدیدی را که به طریقی دربرگیرنده تعامل مناسبی از پارامترهای بهینه‌سازی و معیارهای آمار شکل است، در بخش بعد پیشنهاد نماییم.

۱.۴ الگوریتم گرادیان نزول استوار

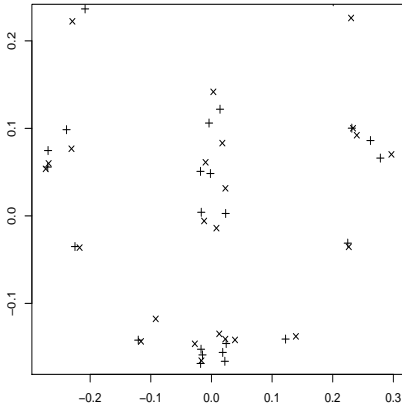
در این بخش برای رفع ضعف بیان شده در بخش قبل، الگوریتم جدیدی پیشنهاد می‌شود. چون در آمار شکل، حفظ هندسه اشکال از اهمیت بسیاری برخوردار است، الگوریتم طوری طراحی می‌شود که تحت تغییر عوامل موثر مانند τ و ε استوار باشد. به همین دلیل الگوریتم گرادیان نزول استوار^۵ (RGDA) نامیده می‌شود.

یکی از مزیت‌های آن این است که در هر مرحله از الگوریتم، میزان انحراف هندسی خروجی از مقدار واقعی، براساس معیار میانگین مجموع توان‌های دوم عادی^۶ (MOPA) بررسی می‌شود. محاسبه این ملاک به این ترتیب است که با استفاده از انطباق بهینه به روش پروکراستس، در هر مرحله از اجرای الگوریتم استاندارد، اختلاف بین ساختار هندسی خروجی الگوریتم و هر یک از ماتریس‌های پیکره‌بندی مشاهدات را محاسبه کرده، سپس میانگین حسابی این اختلاف‌ها به عنوان ملاکی برای ارزیابی میزان انحراف از ساختار هندسی، استفاده شود.

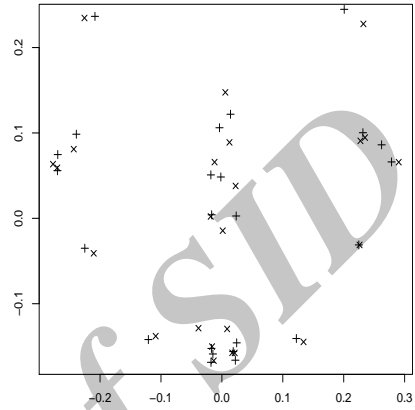
بنابراین با توجه به این ملاک می‌توان الگوریتم مورد مطالعه در این مقاله را به‌گونه‌ای تعمیم داد که هندسه شکل در هر مرحله مورد بازبینی قرار گیرد. سپس

^۵ Robust Gradient Descent Algorithm

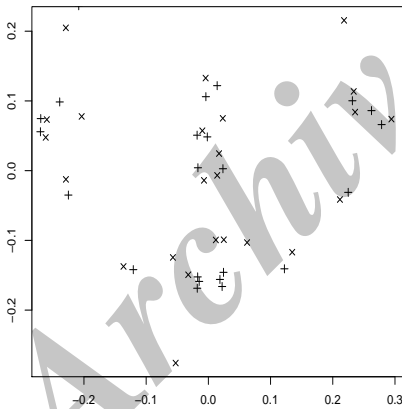
^۶ Mean of Ordinary Procrustes Analysis



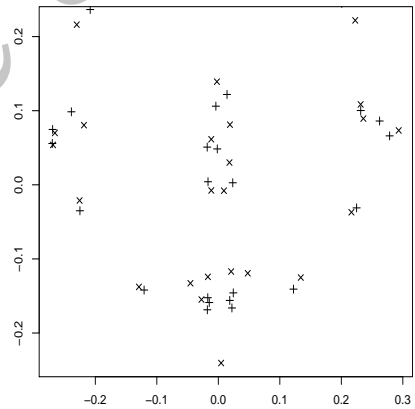
(ب)



(الف)



(د)



(ج)

شکل ۳: میانگین پروکراسس (+) و ذاتی (x) ساختار هندسی مغز از طریق اجرای GDA. الف: $\tau = 0/1$ ، $OPA = 0/0/127$ ، ب: $\tau = 0/4$ ، $OPA = 0/0/561$ ، ج: $OPA = 0/0/561$ ، $\tau = 1$ و د: $OPA = 0/1072$ ، $\tau = 0/7$

براساس حفظ یا از بین رفتن هندسه شکل تعیین شود. به این ترتیب که اگر هندسه شکل در هر مرحله حفظ شد، الگوریتم فرایند عادی مربوط به خود را طی کند، اما اگر هندسه شکل از بین رفت، الگوریتم گرادیان نزول طوری طراحی شود که میانه میانگین‌های به‌دست آمده در مرحله قبلی و اخیر را به‌دست آورده و با در نظر گرفتن آن به‌عنوان نقطه شروع جدید مجدداً فرایند مرحله قبل را طی کند.

۲.۴ ارزیابی عملکرد RGDA

در این بخش با اجرای الگوریتم پیشنهاد شده روی داده‌های مغز و محاسبه میانگین ذاتی شکل به‌ازای مقادیر τ متعلق به بازه $[1, 0.1]$ ، علاوه بر مطالعه برخی ویژگی‌های کلیدی آن، عملکردش در مقابل حالت استاندارد مورد ارزیابی قرار می‌گیرد. روش انجام آزمایش مشابه بخش قبل است.

تعداد تکرارهای لازم تا رسیدن به همگرایی توسط RGDA در جدول ۲ آمده است. نکته قابل تأمل عدم واگرایی این روش به‌ازای کلیه ترکیب‌های در نظر گرفته برای τ و ε است. این مطلب خود به‌نوعی ارجحیت روش پیشنهادی را نشان می‌دهد. به‌علاوه، ملاحظه می‌شود که برای هر مقدار ثابت ε با افزایش τ تعداد تکرار نیز افزایش می‌یابد و یک روند معکوس برای تعداد تکرارها با افزایش ε به‌ازای مقادیر ثابت τ قابل مشاهده است. این نتایج تاییدی بر ادعای مطرح شده در بخش قبل مبنی بر بهبود GDA با اختیار مقادیر بزرگ ε و مقادیر کوچک τ است.

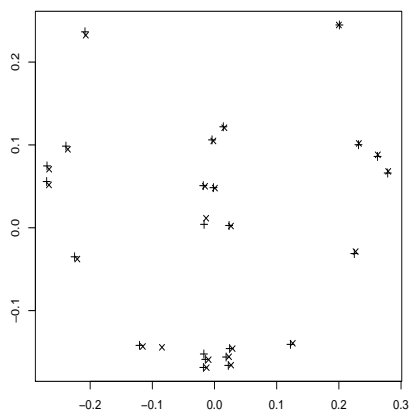
جدول ۲: تعداد تکرار در اجرای روش RGDA تا رسیدن به همگرایی.

| τ | ε | | | |
|--------|---------------|-----|-----|------|
| | ۰/۱ | ۰/۶ | ۰/۱ | ۰/۰۶ |
| ۰/۱ | ۱ | ۱ | ۱ | ۴ |
| ۰/۴ | ۱ | ۱ | ۱ | ۵ |
| ۰/۷ | ۱ | ۲ | ۲ | ۵ |
| ۱ | ۳ | ۵ | ۴ | ۷ |

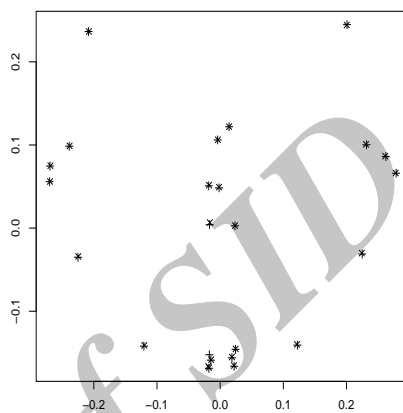
اکنون با ثابت نگه داشتن ε در مقدار $0/06$ به بررسی پایداری ساختار هندسی میانگین ذاتی به ازای مقادیر τ پرداخته می‌شود. نتایج حاصل به صورت نموداری با مختصات دکارتی میانگین ذاتی شکل همراه با کمیت OPA به ازای مقادیر متفاوت τ در شکل ۴ نمایش داده شد. به علاوه مقایسه عینی میانگین پروکراستس نیز در نمودار گنجانده شده است. روند نزولی کمیت OPA برای وقتی که τ افزایش می‌یابد همانند روند مشابه در بخش قبل است. اما نکته حائز اهمیت کاهش این مقادیر برای حالت متناظر آن با استفاده از GDA است. این نتیجه نشان می‌دهد که عملکرد الگوریتم ارائه شده به مراتب بهتر از حالت استاندارد است. در نهایت برای دو مقدار ε برابر $0/6$ و $0/06$ ملاک MOPA به ازای تمامی مقادیر τ در بازه $[0/1, 1]$ ، محاسبه و نتایج حاصل در شکل ۵ نشان داده شده است. ملاحظه می‌شود که با افزایش τ این ملاک نیز افزایش می‌یابد. اما این روند افزایشی برای ε های بزرگ به مراتب کمتر از ε های کوچک است و در سراسر حوزه تغییرات τ ، مقدار آن برای الگوریتم جدید همواره کمتر از استاندارد است. نتیجه کلی این است که با استفاده از RGDA نه تنها معیارهای اختلاف کمتر شده، بلکه پایداری ساختار هندسی برای میانگین ذاتی شکل نیز بیشتر شده است. اضافه بر این با استفاده از آن وجود میانگین ذاتی شکل همواره تضمین می‌شود.

بحث و نتیجه گیری

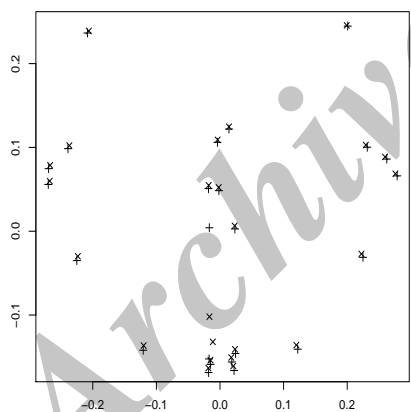
در این مقاله با تعمیم تحلیل ژئودزیک اصلی به فضای شکل، واریانس مولکول DNA محاسبه و با نتایج تحلیل مولفه اصلی مورد مقایسه قرار داده شد. پایداری هندسه میانگین شکل براساس پارامترهای الگوریتم گرادیان نزول برای توسیع تحلیل ژئودزیک اصلی به فضای شکل ضروری است. لذا براساس داده‌های پزشکی مغز انسان، اثر نامطلوب آنها مورد بررسی و مطالعه قرار گرفت. سپس با ارائه ملاک مناسب برای تعیین ساختار هندسی اشیاء، الگوریتم استوارتری نسبت به حالت استاندارد پیشنهاد شد. به علاوه با اجرای الگوریتم پیشنهادی روی داده‌های مغز، ارجحیت آن در بهبود عملکرد تحلیل ژئودزیک اصلی نشان داده شد.



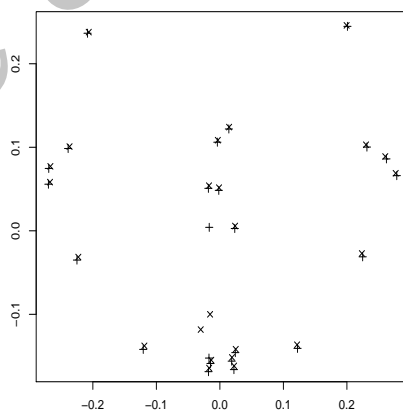
(ب)



(الف)

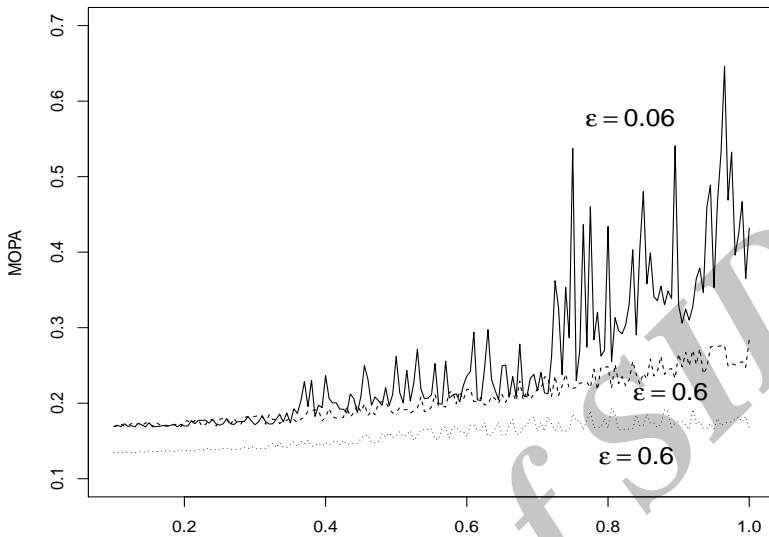


(د)



(ج)

شکل ۴: میانگین پروکراسیس (+) و ذاتی (x) ساختار هندسی مغز از طریق اجرای RGDA. الف: $\tau = 0/1$, $OPA = 0/00021$; ب: $\tau = 0/4$, $OPA = 0/0048$; ج: $OPA = 0/0178$, $\tau = 0/7$ و د: $OPA = 0/0130$, $\tau = 1$.



شکل ۵: مقایسه میزان MOPA حاصـل از RGDA به ازای خطای ثابت ۰/۶ (نقطه چین) با حالت استاندارد به ازای خطاهای ۰/۶ (خط چین) و ۰/۰۶ (خط ممتد) جهت برآورد میانگین ذاتی شکل

تقدیر و تشکر

نویسندگان از داوران محترم که پیشنهادهای سازنده‌ای برای بهبود این مقاله ارائه کردند تقدیر و تشکر می‌نمایند.

مراجع

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

Buss, S. R. and Fillmore, J. P. (2001), Spherical Averages and Application to Spherical Splines and Interpolation, *ACM Transaction on Graphics*, **20**, 95-126.

Cauchy. A. (1847), Methodes Generales Pour La Resolution Des Systemes Dequations Simultanees. *Compte Rendu Academie des Sci-*

ences. Paris, 25, 536-538.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham. J. (1992), Training Models of Shape from Sets of Examples, In Hogg, D. C. and Boyle, R. D. (eds), *British Machine Vision Conference, 9-18*, Springer-Verlag, Berlin.

Dryden, I. and Mardia, K. (1998), *Statistical Shape Analysis*. New York, Wiley.

Dryden, I., Kume, A., Le, H., Wood, A. T. and Laughton, C. (2002), Size-and-Shape Analysis of DNA Molecular Dynamics Simulations, *In Proceedings of Statistics of Large Datasets, LASR2002, University of Leeds, Leeds, 23-26.*

Fotouhi, H. and Golarizadeh, M. (2012). Exploring Variability of DNA Moleculares via Principal Geodesic Analysis on the Shape Space. *Journal of Applied Statistics, 39*, 2199-2207.

Fletcher, P. (2004). *Statistical Variability in Nonlinear Spaces: Application to Shape Analysis and DT-MRI*, Ph.D Thesis, University of North Carolina at Chapel Hill.

Fletcher, P., Lu, C., Pizer, S., Joshi, S. (2004), Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape, *Medical Imaging, IEEE Transactions, 23*, 995-1001.

Free, S. L., O'Higgins, P., Maudgil, D. D., Dryden, I. L., Lemieux, L., Fish, D. R. and Shorvon, S. D. (2001), Landmark-Based Morphometrics of the Normal Adult Brain Using MRI. *Neuroimage, 13*, 801-813.

- Kendall, D. G. (1984), Shape Manifolds, Procrustean Metrics and Complex Projective Spaces, *Bulletin of the London Mathematical Society*, **16**, 81-121.
- Kent, J. T. (1994), The Complex Bingham Distribution and Shape Analysis, *Journal of the Royal Statistical Society. Series B*, **56**, 285-299.
- Luenberger, D. G. (1969), *Optimization by Vector Space Method*. New York, Wiley.
- Mardia, K. and Jupp, P. (2000), *Directional Statistics*, New York, Wiley.
- Meza, J. C. (2010), Steepest Descent, *Wiley Interdisciplinary Review: Computational Statistics*, **2**, 719-722.
- Pennec, X. (1999), Probabilities and Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements, *IEEE Workshop on Non-linear Signal and Image Processing, Antalya, Turkey*.
- Sommer, S., Lauze, F., Hauberg, S. and Nielson, M. (2010), Manifold Valued Statistics, Exact Principal Geodesic Analysis and the Effect of Linear Approximations, *Lecture Notes in Computer Science*, **63**, 43-56.