Journal of Statistical Sciences, Autumn and Winter, 2022
Vol. 16, No. 2, pp 449-468
DOI: 10.29252/jss.16.2.449

Semiparametric Multinomial Logistic Regression Model to Classify Shape Data

Moghimbeygi, M.
Department of Mathematics, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran.

Corresponding author: M. Moghimbeygi, M.Moghimbeygi@khu.ac.ir

Received: 28/8/2022 **Revised:** 22/11/2022 **Accepted and Published Online:** 6/12/2022.

Introduction

Statistical shape analysis is one of the fields of multivariate statistics, where the main focus is on the geometric structures of objects. This analysis method is widely used in many scientific fields, such as medicine and morphology. One of the tools for diagnosing diseases or determining animal species is the images and the shapes extracted from them. Introducing methods of classifying shapes can be a solution to determine the class of each observation. Usually, in regression modelling, explanatory and dependent variables are quantitative. However, one may want to measure the relationship between an explanatory variable (with continuous values) and a dependent variable with qualitative values. One option is to use the multinomial logistic regression model. Therefore, a semiparametric multinomial logistic regression model to classify shape data is introduced in this paper.

Material and Methods

The power-divergence criterion is a measure for hypothesis testing in multinomial data. This criterion is used to define the kernel function of explanatory variables. The model is a multinomial logistic regression model based on kernel function as a function of explanatory variables and an intercept. Since the shapes' geometric structure and size play a key role in the classification of shapes, the kernel function is determined based on the shape distances. The smoothing parameter was estimated using the least square cross-validation method. Also, the estimation of model parameters was done using the neural network method.

Results and Discussion
The shape space is a manifold, but most of the methods presented in the literature for classifying shapes were done in the shape tangent space or used linear transformations. Since mapping from the manifold to linear space decreases data information, applying tangent space and linear spaces will reduce classification accuracy. Therefore, the shape space is used to classify the shape data. The performance of the model in a simulation study and two real data sets were investigated in the paper. The two real data sets used in this paper are taken from the shape package in R software. The first data set is related to schizophrenia patients and people as control, and the second one is associated with the skull of three species of apes of two sexes. The classification of these data showed an accuracy of 82% and 84%, respectively. Also, a comparison was made with the previous methods based on a real data set, which showed the proper performance of our approach compared to the other two techniques.

Conclusion
Since in the nonparametric kernel function, suitable distances of the shape space were used, the introduced method performs better than those based on Euclidean spaces. Also, the ability to use other shape distances, such as partial, full Procrustes and Riemannian distances, makes the model more flexible in classifying different types of shape data. On the other hand, size-and-shape distance can be used in the kernel function to classify data whose size plays a key role in their geometric structure. Furthermore, since few statistical distributions have been introduced in the shape space, nonparametric methods can be helpful in the analysis of shape data. However, using nonparametric methods in the shape space is time-consuming from the point of view of computer calculations.

Keywords: Logistic regresion, Semiparametric regression, Shape data, Classification.

Mathematics Subject Classification (2010): 57N25, 62G08.

مجله علوم آماری، پاییز و زمستان ۱۴۰۱
جلد ۱۶، شماره ۲، ص ۴۴۹ – ۴۶۸
DOI: 10.29252/jss.16.2.449
مقاله پژوهشی

رگرسیون لوژستیک چندجمله‌ای نیمه پارامتری برای رده‌بندی داده‌های شکل

میثم مقیم‌بیگی
گروه ریاضی، دانشکده علوم ریاضی و کامپیوتر، دانشگاه خوارزمی

نویسنده مسئول : میثم مقیم بیگی، M.Moghimbeygi@khu.ac.ir

تاریخ دریافت: ۱۴۰۱/۶/۲۷ تاریخ بازنگری: ۱۴۰۱/۹/۷ تاریخ پذیرش و انتشار: ۱۴۰۱/۹/۱۵

چکیده: در این مقاله یک مدل رگرسیون لوژستیک چند جمله‌ای نیمه پارامتری برای رده‌بندی پیکربندی‌های برچسب‌دار معرفی شده است. در مدل رگرسیونی متغیر تبینی تابع هسته‌ای است که با استفاده از معیار توان–واگرایی به‌دست آمده است. همچنین متغیر پاسخ به‌صورت رسته‌ای بوده و رده هر پیکربندی را نشان می‌دهد. این مدل رگرسیونی نیمه پارامتری بر اساس فواصل تعریف شده در فضای شکل معرفی شده و به همین دلیل میزان رده‌بندی درست اشکال با استفاده از این روش در مقایسه با روش‌های پیشین بهبود یافته است. عملکرد این مدل در قالب یک مطالعه شبیه‌سازی مورد بررسی قرار گرفته است. در انتها نیز کاربردی از این روش در رده‌بندی دو مجموعه داده واقعی به نمایش گذاشته شد. همچنین روش ارائه شده در این مقاله با روش‌های معرفی شده در نوشتگان مقایسه گردید که نشان از عملکرد مناسب این روش در رده‌بندی پیکربندی‌ها دارد.

واژه‌های کلیدی: رگرسیون لوژستیک، رگرسیون نیمه پارامتری، داده شکل، رده‌بندی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62G08 ،57N25.

۱ مقدمه

تحلیل آماری شکل یکی از حوزه‌های آمار چند متغیره است که در آن تمرکز اصلی بر ساختارهای هندسی اجسام قرار دارد. ”شکل یک شیء تمام اطلاعات هندسی باقی مانده از یک شیء پس از حذف اثرات مکان، مقیاس و دوران است.“ این تعریف رسمی برای اولین بار توسط کندال (۱۹۷۷) در مورد شکل ارائه شد. پس از تعریف

شکل سؤالی که مطرح می‌شود این است که "چگونه می‌توان اطلاعات هندسی یک شئ را به‌صورت کمی بیان کرد؟" برای جواب به این سؤال باید گفت که یکی از روش‌های ثبت شئ استفاده از ایده ریخت‌شناسان است. آن‌ها یک شئ را با کمک نقاطی که بر سطح خارجی آن مشخص کرده‌اند معرفی می‌کنند. انتخاب این نقاط باید به‌گونه‌ای باشد که تا حد امکان اطلاعات کلی موجود در شئ را بیان کند. واضح است که هرچه تعداد این نقاط بیشتر باشد، اطلاعات بیشتری از شئ در دسترس خواهد بود. در تحلیل آماری شکل به این نقاط کلیدی نقاط شاخص[1] می‌گویند. به مجموعه تمام نقاط شاخصی که برای یک شئ در نظر گرفته می‌شود، پیکربندی[2] آن شئ گویند (درایدن و ماردیا ، ۲۰۱۶). چنانچه به نقاط شاخص نام یا برچسب اختصاص یابد در این‌صورت به آن‌ها پیکربندی‌های برچسب‌دار گفته می‌شود. به‌منظور تحلیل آماری شکل، ابتدا لازم است پیکربندی‌ها ثبت شوند. دو رویکرد اصلی در ثبت پیکربندی‌ها استفاده از نقاط شاخص و منحنی‌ها است. در این مقاله رویکرد استفاده از نقاط شاخص دنبال می‌شود. روش‌های مختلفی برای ثبت اشکال براساس نقاط شاخص وجود دارد. برای مثال ثبت نقاط با استفاده از زوایا یا اعداد مختلط را می‌توان نام برد (درایدن و ماردیا ، ۲۰۱۶). دو سیستم مختصات شناخته شده برای تحلیل آماری اشکال توسط کندال (۱۹۸۴) و بوکشتین (۱۹۸۶) معرفی شده‌اند. در حالی که مختصات اول عمدتاً برای تحلیل‌های نظری مورد استفاده قرار می‌گیرد، مختصات دوم ابزار مهمی در تحلیل اکتشافی است. منابع برای آشنایی با تحلیل آماری شکل درایدن و ماردیا (۲۰۱۶) و بسته shape در نرم افزار R است (درایدن ، ۲۰۱۲).

برای خلاصه کردن داده‌ها رویکردهای متفاوتی در آمار وجود دارد. یکی از این روش‌ها، رده‌بندی داده‌ها بر اساس یک یا چند ویژگی آن‌ها است. برای رده‌بندی اشکال نیز روش‌های مختلفی معرفی شده است. برای مثال، دباولار و همکاران (۲۰۲۰) خوشه‌بندی داده‌های شکل که در طول زمان تغییر می‌کردند را در رده‌های مختلف بدون برچسب با استفاده از یک مدل ترکیبی جدید مطالعه کردند. مدل آن‌ها یک مسیر با استفاده از میانگین اشکال برای هر خوشه ایجاد می‌کند که از این منحنی‌ها برای خوشه‌بندی استفاده می‌شود. از نقطه نظر کندال (۱۹۷۷) شکل یک شئ معادل نقطه‌ای روی یک کره است. لازم به ذکر است که منفیلد فضای توپولوژی است که در هر نقطه به‌صورت موضعی شبیه فضای اقلیدسی است. با توجه به این‌که کره یک منیفلد است، یکی دیگر از رویکردها، خوشه‌بندی بر روی منیفلد است. این رویکرد برای خوشه‌بندی اشکال، مشابه استفاده از الگوریتم‌های تصویر غیرخطی (یانکف و کئوق ، ۲۰۰۶). نمونه دیگری از این رویکرد، تصویر اشکال در فضای خطی و استفاده از مولفه‌های اصلی برای رده‌بندی اشکال با استفاده از مدل رگرسیون لوژستیک است (مقیم بیگی و نودهی ، ۲۰۲۲).

رگرسیون لوژستیک رابطه بین متغیرهای توضیحی و یک متغیر پاسخ دو وجهی را توصیف می‌کند. این مدل‌بندی ابزار مناسبی برای رده‌بندی داده‌ها با پاسخ‌های دو وجهی است. تعمیم این روش را می‌توان در قالب پاسخ‌های چند وجهی به‌صورت رگرسیون لوژستیک چند جمله‌ای تعریف کرد. با توجه به محدودیت‌های موجود در مسائل کاربردی، محققان مدل رگرسیون لوژستیک را برای غلبه بر آن محدودیت‌ها تعمیم داده‌اند. برای مثال، لین و چن (۲۰۰۸) مدل‌های رگرسیون لوژستیک با پاسخ‌های ترتیبی را با استفاده از مدل‌های لوجیت تجمعی تعمیم دادند. آن‌ها همچنین

[1] Landmarks
[2] Configuration

یک روش هموارسازی خطی موضعی ناپارامتری را برای برازش به پاسخ‌های ترتیبی با استفاده از متغیرهای کمکی و طبقه‌ای معرفی کردند. رده‌بندی بیماران دیابتی استفاده کردند. به‌عنوان مثال دیگری از کاربردهای این روش ریفادا و همکاران (۲۰۲۱) مدل‌های پارامتری را به مدل‌های ناپارامتری برای استفاده در رگرسیون لوژستیک تعمیم دادند. سیمو و همکاران (۲۰۲۰) مدل‌های نیمه خطی تعمیم‌یافته با متغیرهای کمکی روی منیفلد ریمانی را معرفی کردند که این مدل خود شامل یک مدل خطی از متغیرهای تبینی، یک تابع از اندازه‌ها روی منیفلد ریمانی و خطاهایی از توزیع نرمال بود. اکومورا و نایتو (۲۰۰۶) روش دیگری برای رده‌بندی داده‌ها معرفی کردند. آن‌ها یک روش هموارسازی هسته برای رگرسیون چند جمله‌ای ارائه کردند. برآوردگر توابع رگرسیون در مدل آن‌ها با به حداقل رساندن معیار توان–واگرایی[1] موضعی به‌دست می‌آید. این معیار کلاسی از آزمون‌های نیکویی برازش را ارائه می‌دهد.

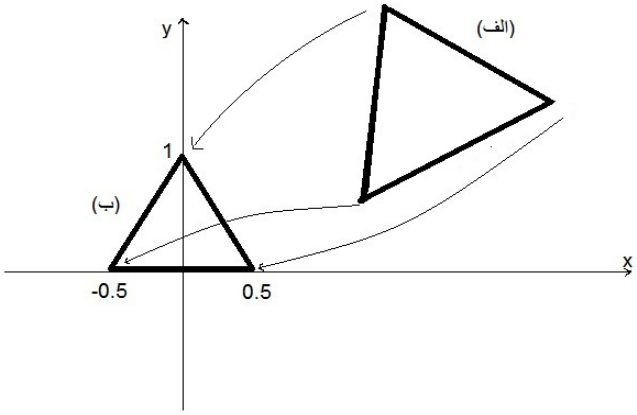
بیشتر روش‌های ارائه شده برای تحلیل اشکال در فضای مماسی شکل انجام گرفته یا از ابزارهای خطی بهره برده‌اند. هرچند مطالعاتی نیز در فضای شکل انجام گرفته است. برای مثال، مقیم بیگی و گل‌علی‌زاده (۱۳۹۸) به مدل‌بندی رگرسیونی داده‌های شکل بر روی کره پرداخته‌اند و فتوحی و گل‌علی‌زاده (۱۳۹۱) عملکرد تحلیل ژئودزیک در فضای شکل را بهبود داده‌اند. همانگونه که اشاره شد فضای شکل یک منیفلد است. لذا بنظر می‌رسد استفاده از فضای مماسی و فضاهای خطی از دقت رده‌بندی بکاهد. از این رو در این مقاله قصد داریم از ابزارهای مناسب فضای شکل برای رده‌بندی اشکال استفاده کنیم. در این مقاله یک مدل رگرسیون لوژستیک چند جمله‌ای[2] (MLR) بر اساس یک برآوردگر ناپارامتری از پیکربندی‌ها ارائه می‌کنیم. در این مدل، پیکربندی‌ها به‌صورت متغیرهای کمکی ظاهر می‌شوند که در آن معیار توان–واگرایی، بخش ناپارامتری مدل را نتیجه می‌دهد. از این رو در بخش ۲، مقدمه‌ای از تحلیل آماری شکل ارائه خواهد شد. سپس، MLR ناپارامتری برای داده‌های شکل با استفاده از معیار توان–واگرایی در بخش ۳ ارائه می‌شود. عملکرد مدل در یک مطالعه شبیه‌سازی و دو مجموعه داده واقعی در بخش ۴ مورد بررسی قرار خواهد گرفت. همچنین مقایسه‌ای نیز با روش‌های پیشین بر اساس یک مجموعه داده واقعی انجام خواهد گرفت. در انتها نیز مقاله با یک بخش نتیجه‌گیری به پایان می‌رسد.

۲ مقدمه‌ای از تحلیل آماری شکل

ر تحلیل آماری شکل به مجموعه نقاط محدودی که هندسه یک شیء را تعیین می‌کند، نقاط شاخص و مجموعه تمام نقاط شاخص در نظر گرفته شده برای یک شیء را پیکربندی می‌نامند. علاوه‌بر این، از یک ماتریس $k \times m$ بعدی مانند X برای تعریف پیکربندی استفاده می‌شود که در آن k تعداد نقاط شاخص و m بعد پیکربندی است. طبق تعریف کندال (۱۹۷۷)، "شکل"، تمام اطلاعات هندسی باقیمانده از یک شیء پس از حذف اثرات مکان، مقیاس و دوران است. مثالی از این تعریف در شکل ۱ قابل مشاهده است. بنابراین، برای به‌دست آوردن شکل یک پیکربندی، لازم است اثرات مکان، مقیاس و دوران حذف شود. برای حذف اثر مکان می‌توان از ماتریس‌های مرکزی کننده مانند

[1] Power-divergence
[2] Multinomial logistic regression

زیرماتریس هلمرت[1]، نمایش داده شده با نماد H، استفاده کرد. (درایدن و ماردیا ، ۲۰۱۶). از این رو، ماتریس HX پیکربندی مرکزی یا هلمرتی شده نامیده می‌شود. در ادامه چند معیار فاصله که می‌توانند در تحلیل آماری داده‌های شکل مورد استفاده قرارگیرند معرفی می‌شوند.



شکل ۱. حذف اثر مکان، مقیاس و دوران مثلث متساوی الاضلاع در مختصات بوکشتین با انتقال دو نقطه شاخص مثلث (الف) به $(-۰٫۵, ۰)$ و $(۰٫۵, ۰)$. مثلث (ب) شکل و مثلث (الف) پیکربندی است.

تعریف ۱. فرض کنید $X_۱$ و $X_۲$ دو ماتریس پیکربندی $k \times m$ باشند. برای $j = ۱, ۲$، با فرض $Z_j = HX_j / \parallel HX_j \parallel$, که به‌عنوان قبل-شکل[2] نامیده می‌شود، فاصله ریمانی کوتاه‌ترین فاصله بین $Z_۱$ و $Z_۲$ بر روی کره قبل-شکل است و به‌صورت $\rho(X_۱, X_۲) = \arccos(\sum_{i=۱}^{m} \lambda_i)$، محاسبه می‌شود، که در آن λ_iها مقادیر ویژه حاصل از تجزیه مقدار تکین عبارت زیر است

$$Z_۱^T Z_۲ = V \operatorname{diag}(\lambda_۱, \ldots, \lambda_m) U^T. \qquad (۱)$$

تعریف ۲. فاصله پروکراستس[3] جزئی با انطباق بهینه ماتریس‌های قبل-شکل $Z_۱$ و $Z_۲$ بر اساس دوران حاصل می‌شود و از رابطه $d_p(X_۱, X_۲) = \sqrt{۲}(۱ - \sum_{i=۱}^{m} \lambda_i)^{۱/۲}$ به‌دست می‌آید.

تعریف ۳. فاصله پروکراستس[4] تام که با نماد $d_F(X_۱, X_۲)$ نمایش داده می‌شود با انطباق بهینه ماتریس‌های قبل-شکل $Z_۱$ و $Z_۲$ بر اساس دو عامل مقیاس و دوران به‌دست می‌آید و به‌صورت $d_F(X_۱, X_۲) = [۱ -$

[1] Helmert
[2] pre-shape
[3] Procrustes
[4] Procrustes

$^{1/2}[(\sum_{i=1}^{m}\lambda_i)^2]$، قابل محاسبه است.

تعریف ۴. فاصله ریمانی در فضای اندازه-و-شکل[1] که با نماد $d_S(X_1, X_2)$ نمایش داده می‌شود به‌صورت $d_S(X_1, X_2) = \sqrt{S_1^2 + S_2^2 - 2S_1S_2\cos\rho(X_1, X_2)}$ است، که در آن S_1 و S_2 اندازه‌های مرکزی X_1 و X_2 هستند.

بنابه دریدن و ماردیا (۲۰۱۶)، با استفاده از فاصله ریمانی، می‌توان توزیع اندازه-و-شکل را مستقیماً در فضای اندازه-و-شکل مشخص کرد. یک توزیع اندازه-و-شکل متقارن دورانی[2]، متناسب با عبارت

$$\exp\{-h\phi(d_S(X,\mu))\}, \qquad (۲)$$

است، که در آن $\phi(x)$ یک تابع مثبت است. یک انتخاب مناسب برای این تابع $\phi(x) = x^2$ است که در آن ثابت نرمال‌ساز تابع چگالی، وابسته به پارامتر μ نیست. ملاک RMS معیاری است که در تحلیل آماری شکل برای محاسبه همگنی استفاده می‌شود. این معیار تعدیلی از اندازه d_S است که به‌صورت $\text{RMS} = d_S/\sqrt{k}$ تعریف می‌شود که در آن k تعداد نقاط شاخص است. مقدار کوچک RMS به معنای همگن بودن مجموعه داده شکل است.

تعریف ۵. تجزیه و تحلیل جزئی پروکروستس معمولی[3] (POPA) با استفاده از تبدیلات اقلیدسی جابجایی و دوران دو پیکربندی با حداقل رساندن عبارت

$$\| X_2 - X_1\Gamma - 1_k\gamma^T \|^2, \qquad (۳)$$

تعریف می‌شود. می‌توان نشان داد مقادیر بهینه برای γ و Γ با استفاده از تجزیه مقدار تکین (۱) به‌صورت $\hat{\gamma} = 0$ و $\hat{\Gamma} = UV^T$ به‌دست می‌آید، که در آن U و V ماتریس‌های متعامد ویژه ($SO(m)$) تعریف شده در معادله (۱) هستند (دریدن و ماردیا، ۲۰۱۶). بنابراین انطباق جزئی X_1 بر روی X_2 به صورت $X_1^p = X_1\hat{\Gamma} + 1_k\hat{\gamma}^T$ است.

۳ رگرسیون لوژستیک چندجمله‌ای نیمه پارامتری

رگرسیون نیمه پارامتری در واقع ترکیبی از مدل‌های رگرسیونی پارامتری و ناپارامتری است. این مدل‌ها اغلب زمانی استفاده می‌شوند که مدل ناپارامتری عملکرد خوبی نداشته باشد یا زمانی که محقق می‌خواهد از یک مدل پارامتری استفاده کند اما ساختار داده‌ها به‌گونه‌ای است که مدل‌بندی آن‌ها با استفاده از مدل‌های شناخته شده پارامتری امکان‌پذیر نیست. فرض کنید X_i یک ماتریس پیکربندی و یک متغیر (تبینی) با متغیر پاسخ دو وجهی ($Y_i(\in\{0,1\}$ برای

[1] Size-and-shape
[2] Rotationally symmetric
[3] Partial Ordinary Procrustes Analysis

$i = ۱, ۲, \ldots, n$ باشد. در زیر بخش بعد قصد داریم یک مدل MLR برای رده‌بندی پیکربندی‌ها بر اساس عضویت آن در گروه‌های متناظرشان ارائه کنیم. برای این منظور، ابتدا تبدیل ناپارامتری پیکربندی‌ها را بر اساس معیار توان-واگرایی ارائه می‌کنیم. سپس یک مدل لوژستیک نیمه پارامتری در فضای اندازه-و-شکل معرفی می‌کنیم. مدل معرفی شده توانایی تفکیک پیکربندی‌ها را براساس معیارهای آماری مناسب دارد.

۳.۱ تبدیل ناپارامتری پیکربندی‌ها

برای $i = ۱, \ldots, K$، فرض کنید x_iها متغیرهای تبینی Y_iها متغیرهای پاسخ و $Y_i = (Y_{i۱}, \ldots, Y_{ir})^T$ دارای توزیع چندجمله‌ای $N_i = N(x_i) = \sum_{j=۱}^{r} Y_{ij}$ باشد که در آن $MN(p_۱(x_i), \ldots, p_r(x_i); N(x_i))$ و $Y_{ij} = Y_j(x_i)$. به‌منظور ارزیابی برآوردگرها، روش‌های متفاوتی وجود دارد. یکی از آن‌ها را که می‌توان برای آزمایش‌هایی که از توزیع چندجمله‌ای تبعیت می‌کنند استفاده کرد، آزمون نیکویی برازش است. این آزمون بر اساس معیار توان-واگرایی به‌دست آمده است که توسط کرسی و رید (۱۹۸۴) برای ارزیابی برآوردگر به‌صورت

$$I_\lambda(\boldsymbol{p}_i : \boldsymbol{q}) = \frac{۱}{\lambda(۱+\lambda)} \sum_{j=۱}^{r} p_{ij}\{(\frac{p_{ij}}{q_j})^\lambda - ۱\},$$

بدست آورده شده است، که در آن $\lambda \in \mathbb{R}$، $\boldsymbol{p}_i = (p_{i۱}, \ldots, p_{ir})$ و $\boldsymbol{q} = (q_۱, \ldots, q_r)$ پارامترهای توزیع احتمال هستند، معرفی کردند. مشابه روش معرفی شده توسط اکومورا و نایتو (۲۰۰۶)، ما یک نسخه موضعی از توان-واگرایی را با استفاده از تابع هسته ناپارامتری $K_h(\cdot)$ به‌صورت

$$L_\lambda(\boldsymbol{q}, \gamma) = \frac{۱}{\lambda(۱+\lambda)} \sum_{i=۱}^{K} K_h(x_i - x)\Big[\sum_{j=۱}^{r} p_{ij}\Big\{\Big(\frac{p_{ij}}{q_j}\Big)^\lambda - ۱\Big\} + \gamma\Big(۱ - \sum_{j=۱}^{r} q_j\Big)\Big],$$

در نظر می گیریم که γ ضریب لاگرانژ است. پس برآوردگر $q_j(x)$ با به حداقل رساندن $L_\lambda(\boldsymbol{q}, \gamma)$ به‌صورت

$$\hat{q}_{j\lambda}(x; h) = \frac{\{\sum_{i=۱}^{K} K_h(x_i - x) p_{ij}^{\lambda+۱}\}^{۱/(\lambda+۱)}}{\sum_{j=۱}^{r}\{\sum_{i=۱}^{K} K_h(x_i - x) p_{ij}^{\lambda+۱}\}^{۱/(\lambda+۱)}}, \qquad (۴)$$

به‌دست می‌آید. به‌وضوح $\sum_{j=۱}^{r} \hat{q}_{j,\lambda}(x; h) = ۱$ و به‌ازای هر x تساوی $۰ \leq \hat{q}_{j,\lambda}(x; h) \leq ۱$ برقرار است. توجه شود که اگر $\lambda = ۰$، آنگاه برآوردگر حاصل همان برآوردگر نادارایا-واتسن[1] است که به‌صورت

$$\hat{q}_{j,۰}(x; h) = \frac{\sum_{i=۱}^{K} K_h(x_i - x) p_{ij}}{\sum_{i=۱}^{K} K_h(x_i - x)},$$

[1] Nadaraya-Watson

نوشته می‌شود. در این مقاله از برآوردگر (۴) به‌ازای $\lambda = 0$ برای رده‌بندی پیکربندی‌ها استفاده می‌شود.

برای استفاده از مدل معرفی شده در داده‌های شکل، ما r دسته از پیکربندی‌های X_{ij} را برای $1 \le j \le r$ و $1 \le i \le K$ در نظر می‌گیریم. همچنین برای تعریف معیار خود از آماره آزمون گودالز[1] استفاده می کنیم. این آماره به‌منظور آزمون برابری میانگین اشکال در دو جامعه مورد استفاده قرار می‌گیرد. از بخش ۹ از درایدن و ماردیا (۲۰۱۶)، با فرض X_{ij} به‌عنوان یک متغیر تصادفی $k \times m$ از $N_{km}(\mu_{ij}, \sigma^2 I_{km})$، فاصله پروکروستس تام X_{ij} از هر پیکربندی ثابت μ_j برابر $d_F^2(X_{ij}, \mu_j)$ است که تقریباً مشابه χ^2 با درجه آزادی $km - m(m+1)/2 - 1$ توزیع شده است. برای r دسته از پیکربندی‌ها،

$$B_{ij}(X_{ij}, \mu_j) = \frac{d_F^2(X_{ij}, \mu_j)}{\sum_{j=1}^{r} d_F^2(X_{ij}, \mu_j)}$$

یک متغیر تصادفی از توزیع بتا است. چون فاصله هر شکل با اعضای هر رده می‌تواند نقش مهمی در تعیین رده آن شکل داشته باشد لذا در این مقاله از فواصل معرفی شده در بخش ۲ برای رده‌بندی اشکال استفاده شده است. به‌عبارت دیگر $B_{ij}(X_{ij}, X)$های بدست آمده از پیکربندی X و سایر پیکربندی‌ها نقش اصلی را در رده‌بندی بازی می‌کنند. برآوردگر q_j برای داده‌های شکل را می‌توان با جایگذاری $B_{ij}(X_{ij}, X)$ به‌جای p_{ij} در (۴) به‌صورت

$$\hat{q}_{j\lambda}(X; h) = \frac{\{\sum_{i=1}^{K} K_h(X_{ij} - X) d_F^{2\lambda+2}(X_{ij}, X)\}^{1/(\lambda+1)}}{\sum_{j=1}^{r} \{\sum_{i=1}^{K} K_h(X_{ij} - X) d_F^{2\lambda+2}(X_{ij}, X)\}^{1/(\lambda+1)}}, \qquad (۵)$$

نوشت. از آنجایی که مقدار کوچک $d_F^2(X_{ij}, X)$ به معنای این است که دو پیکربندی X_{ij} و X در یک رده یکسان هستند، بنابراین $\hat{q}_{j\lambda}(X; h)$ احتمال عدم تعلق X به رده j تعریف می‌شود. برای اشکالی که به هم شبیه هستند (فاصله ریمانی کمتری دارند)، داریم $d_p = d_F + O(d_F^3)$ و $\rho = d_F + O(d_F^3)$ (درایدن و ماردیا ، ۲۰۱۶).

در برخی از مسائل واقعی اندازه پیکره‌بندی‌ها نقش مهمی در رده‌بندی آن‌ها دارد. برای مثال جمجمه یک انسان بالغ از نظر شکلی مشابه جمجمه یک کودک است و تنها تفاوت می‌تواند اندازه جمجمه‌ها باشد. لذا در این مسائل استفاده از d_p که اندازه را نیز لحاظ می‌کند، می‌تواند معیار مناسب‌تری در مدل معرفی‌شده باشد. از این رو، d_p می‌تواند به‌جای d_F در معادله (۵) استفاده شود. برای انتخاب تابع هسته، از توزیع (۲) با تابع $\phi(x) = x^2$ استفاده می‌کنیم. انتخاب پهنای باند در تحلیل ناپارامتری از اهمیت بالایی برخوردار است. راه‌کارهای زیادی برای انتخاب پهنای باند وجود دارد. برای مثال می‌توان به روش اعتبارسنجی متقابل حداقل مربعات و اعتبارسنجی متقابل حداکثر درستنمایی اشاره کرد (دوین ، ۱۹۷۶). از آنجایی که هدف اصلی این مقاله ارائه روشی برای رده‌بندی پیکربندی‌ها است، پهنای باند برآورد شده باید بتواند حداکثر رده‌بندی صحیح را انجام دهد. بنابراین، برای برآورد پارامتر پهنای باند از روش اعتبارسنجی متقابل حداقل مربعات استفاده می‌شود تا فاصله بین احتمال برآورد شده و متغیر پاسخ

[1] Goodalls

کمترین مقدار را داشته باشد. اصلاحی از این معیار به‌صورت

$$CV(h) = \sum_{j=1}^{r}\sum_{i=1}^{K}(y_{ij} - q_{j,\lambda}(-X_{ij}, h))^2,$$

معرفی می‌شود که $q_{j,\lambda}(-X_{ij}, h)$ برآوردگری است که در آن مشاهده X_{ij} در هر مرحله کنار گذاشته شده است (واسرمن ، ۲۰۰۶؛ ریفادا و همکاران ، ۲۰۲۱). برای انتخاب پهنای باند بهینه براساس این معیار لازم است هدف محقق مبتنی بر به حداقل رساندن $CV(h)$ باشد. از آنجایی که در تحلیل‌های ناپارامتری معمولاً برآورد پهنای باند فرم بسته‌ای بر اساس حداقل کردن $CV(h)$ ندارد، لذا برآورد h معمولاً با استفاده از الگوریتم‌های بهینه‌سازی و به‌صورت عددی انجام می‌گیرد. از آنجا که B_{ij}ها با فرض این‌که پیکربندی از توزیع چند متغیره نرمال به‌دست می‌آید، انتظار می‌رود که برآوردگر معرفی شده (۵) یک برآوردگر اریب باشد. این مقاله از رگرسیون لوژستیک خطی ساده برای حذف این سوگیری‌ها استفاده می‌کند.

۳.۲ مدل لوژستیک

در مدل‌بندی رگرسیون، مدلی برای پیش‌بینی متغیر وابسته بر اساس متغیر تبینی ایجاد می‌شود. معمولاً در مدل ایجاد شده، هم متغیرهای تبینی و هم متغیرهای وابسته کمی هستند. با این حال، ممکن است بخواهیم رابطه بین یک متغیر تبینی (با مقادیر پیوسته) و یک متغیر وابسته با مقادیر کیفی را اندازه‌گیری کنیم. روش استاندارد رگرسیون خطی در این مورد کارا نخواهد بود که در این موارد می‌توان از MLR استفاده کرد. MLR یک روش برای رده‌بندی داده‌ها به دو یا چند رده مختلف است. در رگرسیون خطی استاندارد، مقدار مورد انتظار متغیر پاسخ Y ترکیبی خطی از متغیرهای تبینی است. یک نسخه نیمه پارامتری از مدل خطی ساده را می‌توان به‌صورت $\mathbb{E}(Y|X = x) = \alpha + \beta q(x)$ تعریف کرد که α پارامتر عرض از مبداء و $q(x)$ یک تابع ریاضی است. هنگامی که برآمد Y دو وجهی است، یک انتخاب استفاده از رگرسیون لوژستیک است. گزینه‌های دیگر برای تحلیل این مدل‌ها را می‌توان در کاکس و اسنیل (۱۹۸۹) یافت. با استفاده از توزیع لوژستیک، می‌توان میانگین Y به شرط x را به‌صورت

$$\pi(x) = \mathbb{E}(Y|x) = \frac{e^{\alpha+\beta q(x)}}{1 + e^{\alpha+\beta q(x)}},$$

تعریف کرد که در آن $\pi(x)$ به‌عنوان احتمال موفقیت نیز شناخته می‌شود. یک تبدیل ساده از $\pi(x)$ که نقش کلیدی در این مقاله ایفا می‌کند تبدیل لوجیت $\ln[\frac{\pi(x)}{1-\pi(x)}] = \alpha + \beta q(x)$ است. برای محاسبه MLR، ابتدا p متغیر کمکی با r رده متفاوت (با پاسخ‌های اسمی Y) و بردار $\boldsymbol{x}$ از متغیرهای تبینی به طول $p+1$ را در نظر بگیرید. با

انتخاب r-امین رده به‌عنوان یک محور داریم

$$g_m(\boldsymbol{x}) = \ln \frac{Pr(Y=m|\boldsymbol{x})}{Pr(Y=r|\boldsymbol{x})} = \boldsymbol{x}^T\boldsymbol{\beta}_m, \quad m = ۱, \ldots, r-۱,$$

که در آن $\boldsymbol{x} = (۱, x_۱, \ldots, x_p)$ برداری به‌طول $p+۱$ و $\boldsymbol{\beta}_m^T = (\beta_{m۰}, \beta_{m۱}, \ldots, \beta_{mp})$ بردار پارامترها است. تعمیمی از احتمال شرطی در r رده ممکن به‌صورت

$$\pi_j(\boldsymbol{x}) = Pr(Y=j|\boldsymbol{x}) = \frac{e^{g_j(\boldsymbol{x})}}{\sum_{k=۰}^{r} e^{g_k(\boldsymbol{x})}}, \quad j = ۰, \ldots, r-۱,$$

است. برای تابع درستنمایی، از r متغیر دو وجهی U_i که مقادیر ۰ یا ۱ را می‌گیرند استفاده می‌شود. بدین صورت که اگر $Y=i$، آن‌گاه $U_i = ۱$ و برای $j \neq i$، $U_i = ۰$ باشد. در این‌صورت بدون در نظر گرفتن مقدار U_i داریم $\sum_{k=۱}^{r} U_k = ۱$. بنابراین تابع درستنمایی برای n نمونه مستقل از مشاهدات به‌صورت

$$L(\boldsymbol{\beta}) = \prod_{i=۱}^{n}\prod_{k=۱}^{r} \pi_k(\boldsymbol{x}_i)^{u_{ki}},$$

است. با لگاریتم‌گیری از تابع درستنمایی و در نظر گرفتن شرط $\sum_{k=۱}^{r} U_k = ۱$، داریم:

$$\ell(\boldsymbol{\beta}) = \sum_{i=۱}^{n}\sum_{k=۱}^{r}[u_{ki}\pi_k(\boldsymbol{x}_i)] - \sum_{i=۱}^{n} \ln\big[\sum_{k=۱}^{r} e^{g_k(\boldsymbol{x}_i)}\big].$$

با مشتق‌گیری از تابع $\ell(\boldsymbol{\beta})$ نسبت به پارامترهای مدل، برای $k = ۱, \ldots r$ و $l = ۰, \ldots, p$ داریم

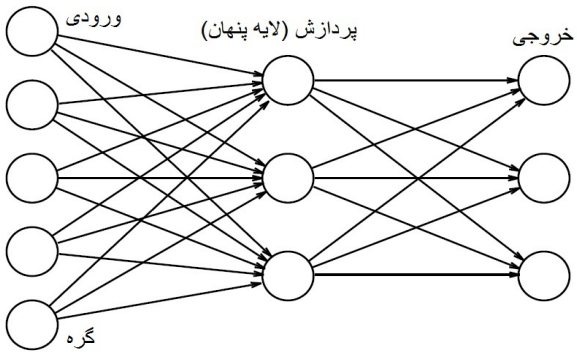
$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_{kl}} = \sum_{i=۱}^{n} x_{li}(u_{ki} - \pi_k(x_i)),$$

که در آن $x_{۰i} = ۱$. چون حل تحلیلی معادله $\partial\ell(\boldsymbol{\beta})/\partial\beta_{kl} = ۰$ میسر نیست یا بسیار دشوار است، باید به روش‌های عددی متوصل شد. شبکه عصبی مصنوعی در واقع یک مدل محاسباتی است که از روش‌های محاسباتی جدید برای یادگیری ماشینی، نمایش دانش و در انتها اعمال دانش به دست آمده در جهت پیش‌بینی پاسخ‌های خروجی از سامانه‌های پیچیده استفاده می‌کند. هر شبکه عصبی از سه لایه تشکیل شده است: ورودی، خروجی و پردازش. هر لایه شامل گروهی از سلول‌های عصبی است که در حالت کلی این سلول‌ها با تمام نورون‌های لایه‌های دیگر در ارتباط هستند. یکی از شبکه‌های عصبی که در مدل‌های رگرسیونی مورد استفاده قرار می‌گیرد شبکه عصبی پیش‌خور[۱]

[1] Feedforward

است. شبکه عصبی پیشخور اولین و ساده‌ترین نوع شبکه عصبی است. در این شبکه اطلاعات تنها از در یک مسیر حرکت می‌کنند. به‌عبارت دیگر اطلاعات با شروع از گره‌های ورودی و گذر از لایه‌های پنهان به سمت گره‌های خروجی می‌روند. نمایی از یک شبکه عصبی پیش‌خور در شکل ۲ قابل مشاهده است.

در شبکه‌های عصبی از یک تابع هزینه برای برآورد پارامترها استفاده می‌شود. در مدل‌های رگرسیونی معمولاً منفی لگاریتم تابع درستنمایی را به‌عنوان تابع هزینه در نظر می‌گیرند. سپس با استفاده از روش گرادیان کاهشی به برآورد پارامترها می‌پردازند. به این معنی که گرادیان تابع را حساب کرده، کمی در خلاف جهت آن حرکت کرده و این کار را آنقدر ادامه داده تا تابع هزینه خیلی کوچک شود.



شکل ۲. شبکه عصبی پیش‌خور همراه با سه لایه ورودی، پنهان و خروجی به‌ترتیب دارای ۵، ۳ و ۳ گره.

برای $i = ۱, \ldots, K$ و $j = ۱, \ldots, r$، فرض کنید ماتریس‌های پیکربندی X_{ij} به‌عنوان متغیرهای تبینی متناظر با پاسخ‌های اسمی y_{ij} در r رده مختلف باشند. برای یک مدل لگ-خطی چند جمله‌ای با r رده، یک شبکه عصبی با r خروجی، منفی لگاریتم تابع درستنمایی شرطی به‌صورت

$$E = -\sum_{i=۱}^{K}\sum_{j=۱}^{r} t_{ij}\log \pi_{ij}, \quad \pi_{ij} = \frac{e^{y_{ij}}}{\sum_{j=۱}^{r} e^{y_{ij}}}$$

است، که در آن t_{ij} هدف و y_{ij} خروجی برای ورودی ij-ام است. یک منبع مناسب برای آشنایی با شبکه‌های عصبی ونابلز و ریپلی (۲۰۰۲) است. همچنین نحوه برآورد پارامترها در مدل MLR با استفاده از شبکه عصبی در نرم افزار R، در این منبع موجود است. در این مقاله از بسته `nnet` برای برازش MLR به‌منظور رده‌بندی پیکربندی‌ها استفاده می‌شود. پس از مدل‌بندی رگرسیونی داده‌ها با استفاده از مدل MLR لازم است عملکرد این مدل رگرسیونی ارزیابی شود. لذا در ادامه چند معیار شناخته شده که در ارزیابی مدل‌های MLR به‌کار می‌رود، معرفی می‌شود. اولین

ملاک، معیار اطلاع آکائیک[1] (AIC) است. این معیار تخمین‌گر خطای پیش‌بینی و در نتیجه کیفیت نسبی مدل‌های آماری برای مجموعه‌ای از داده‌ها است (آکائیک ، ۱۹۷۳). با توجه به مجموعه‌ای از مدل‌های کاندید برای داده‌ها، AIC کیفیت هر مدل را نسبت به هر یک از مدل‌های دیگر تخمین می‌زند. این معیار برای انتخاب مدل به‌صورت

$$AIC = -۲\ell(\hat{\boldsymbol{\beta}}) + ۲(p+۱)$$

تعریف می‌شود. مدل ترجیحی مدلی است که دارای حداقل مقدار AIC باشد. توسیعی از معیار AIC به‌نام معیار اطلاع بیزی[2] (BIC) توسط شوارز (۱۹۷۸) به‌صورت

$$BIC = -۲\ell(\hat{\boldsymbol{\beta}}) + \ln(n)(p+۱)$$

ارائه شد. معیار دیگری که برای ارزیابی مدل‌ها مورد استفاده قرار می‌گیرد، معیار ضریب تعیین است. ضریب تعیین در واقع نسبت تغییرات در متغیر وابسته است که قابل پیش‌بینی از طریق متغیر(های) مستقل است. این ضریب با $R^۲$ نشان داده می‌شود و معمولاً در رگرسیون خطی از آن استفاده می‌شود. در رگرسیون لوژستیک ضرایب تعیین‌های متنوعی معرفی شده است که به‌عنوان مثال می‌توان به ضرایب تعیین مک‌فادن[3] (MF)، کاکس‌اسنیل[4] (CS) و ناگلکرک[5] (Na) اشاره کرد، که به ترتیب به‌صورت

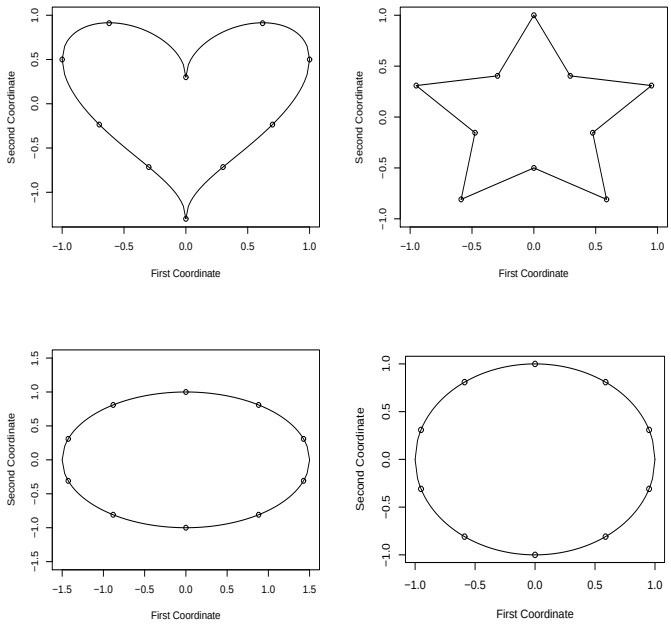
$$R^۲_{MF} = ۱ - \frac{\ell(\hat{\boldsymbol{\beta}})_k}{\ell(\hat{\boldsymbol{\beta}})_{null}}, \quad R^۲_{CS} = ۱ - \left[\frac{\ell(\hat{\boldsymbol{\beta}})_{null}}{\ell(\hat{\boldsymbol{\beta}})_k}\right]^{۲/n}, \quad R^۲_{Na} = \frac{۱ - \left[\frac{\ell(\hat{\boldsymbol{\beta}})_{null}}{\ell(\hat{\boldsymbol{\beta}})_k}\right]^{۲/n}}{۱ - (\ell(\hat{\boldsymbol{\beta}})_{null})^{۲/n}},$$

تعریف می‌شوند. در این ضرایب $\ell(\hat{\boldsymbol{\beta}})_{null}$ مقدار تابع درستنمایی تحت فرض $H_۰ : \beta_{ij} = ۰$ برای $i = ۱, \ldots r-۱$ و $j = ۱, \ldots p$ و $\ell(\hat{\boldsymbol{\beta}})_{Full}$ مقدار تابع درستنمایی تحت فرض تمامی فضای پارامتری است.

۴ مطالعه شبیه‌سازی و تحلیل داده‌های واقعی

در این بخش قصد داریم داده‌های شبیه‌سازی شده و واقعی را رده‌بندی کنیم. لذا در زیربخش اول، داده‌ها را شبیه‌سازی کرده، بر اساس مدل MLR رده‌بندی و عملکرد روش معرفی شده را ارزیابی می‌کنیم. در بخش دوم نیز دو مجموعه داده واقعی که مرتبط با ماهیت پیکربندی (در مفهوم شکل) هستند را رده‌بندی می‌کنیم.

[1] Akaike information criterion
[2] Bayesian information criterion
[3] McFadden
[4] CoxSnell
[5] Nagelkerke



شکل ۳. نمایی از شکل‌های مرجع ستاره، قلب، دایره و بیضی استفاده شده در مطالعه شبیه‌سازی

۴.۱ مطالعه شبیه‌سازی

به‌منظور مطالعه شبیه‌سازی، از روشی که در نبیل و گل‌علی‌زاده (۲۰۱۶) توضیح داده شده است برای شبیه‌سازی پیکربندی‌ها استفاده می‌کنیم. مدل به‌صورت

$$X_{ij} = \beta_{ij}\mu_j\Gamma_{ij} + 1_k\gamma_{ij}^T + \varepsilon_{ij} \quad i = 1\ldots, K, \quad j = 1, \ldots, r, \qquad (۶)$$

تعریف می‌شود که در آن μ_j، j-امین پیکربندی مرجع و ε_{ij} یک ماتریس تصادفی $k \times m$ به‌عنوان منبع خطا است که از توزیع نرمال چند متغیره تولید می‌شود. ما از چهار پیکربندی مرجع، ستاره، قلب، دایره و بیضی با ده نقطه شاخص در دو بعد استفاده کردیم. نمایشی از اشکال مرجع در شکل ۳ نشان داده شده است. ابتدا ۱۰، ۳۰ یا ۵۰ پیکربندی از مدل (۶) برای هر یک از شکل‌های مرجع شبیه‌سازی می‌کنیم. به‌عنوان مثال در حالت ۱۰ پیکربندی، مجموعاً ۴۰ پیکربندی داریم که از مدل (۶) با ۴ میانگین مختلف شبیه‌سازی شده‌اند. سپس با استفاده از روش معرفی شده با استفاده از پهنای باند برآورد شده به رده‌بندی داده‌ها می‌پردازیم. این عمل را ۱۰۰ بار تکرار می‌کنیم.

جدول ۱. میانگین و انحراف معیار ضریب تعیین‌نما (مک‌فادن، کاکس‌اسنیل و ناگلکرک) و معیارهای برازش مدل (AIC، BIC و OCP) بر اساس ۱۰۰ بار تکرار از مدل (۶) برای مقادیر مختلف σ.

تعداد پیکربندی	σ	برآورد	ضریب تعیین‌نما			معیارهای برازش مدل		
			مک‌فادن	کاکس‌اسنیل	ناگلکرک	AIC	BIC	OCP
۱۰	۰٫۱	میانگین	۰٫۹۹۳۹	۰٫۹۳۶۴	۰٫۹۹۸۸	۲۵٫۳۵۹	۶۳٫۴۷۱	۰٫۹۹۶۲
		انحراف معیار	۰٫۰۱۳۰	۰٫۰۰۲۴	۰٫۰۰۲۵	۲٫۸۹۰۳	۲٫۸۹۰۳	۰٫۰۰۸۴
	۰٫۳	میانگین	۰٫۸۶۷۲	۰٫۹۰۷۵	۰٫۹۶۸۰	۳۸٫۷۲۸	۶۵٫۷۵۰	۰٫۹۱۷۵
		انحراف معیار	۰٫۰۸۴۱	۰٫۰۲۰۱	۰٫۰۲۱۵	۹٫۳۲۵۷	۹٫۳۲۵۷	۰٫۰۵۱۴
	۰٫۵	میانگین	۰٫۷۰۶۰	۰٫۸۵۷۱	۰٫۹۱۴۲	۸۹٫۲۲۲	۱۲۷٫۳۳	۰٫۸۰۶۲
		انحراف معیار	۰٫۰۵۹۲	۰٫۰۲۳۳	۰٫۰۲۴۸	۱۳٫۱۲۱	۱۳٫۱۲۱	۰٫۰۵۲۵
۳۰	۰٫۱	میانگین	۰٫۹۹۷۶	۰٫۹۳۷۱	۰٫۹۹۹۵	۲۴٫۷۹۳	۶۹٫۳۹۳	۰٫۹۹۸۳
		انحراف معیار	۰٫۰۰۷۵	۰٫۰۰۱۳	۰٫۰۰۱۴	۲٫۵۰۲۱	۲٫۵۰۲۱	۰٫۰۰۵۳
	۰٫۳	میانگین	۰٫۹۱۷۱	۰٫۹۲۱۰	۰٫۹۸۲۴	۵۱٫۵۸۳	۹۶٫۱۸۳	۰٫۹۵۱۳
		انحراف معیار	۰٫۰۳۷۰	۰٫۰۰۷۹	۰٫۰۰۸۴	۱۲٫۳۱۰	۱۲٫۳۱۰	۰٫۰۲۰۸
	۰٫۵	میانگین	۰٫۷۱۵۵	۰٫۸۶۲۱	۰٫۹۱۹۶	۱۱۸٫۶۶	۱۶۳٫۲۶	۰٫۸۲۱۷
		انحراف معیار	۰٫۰۲۷۶	۰٫۰۱۰۴	۰٫۰۱۱۰	۹٫۱۶۶۵	۹٫۱۶۶۵	۰٫۰۳۵۰
۵۰	۰٫۱	میانگین	۰٫۹۹۸۹	۰٫۹۳۷۳	۰٫۹۹۹۸	۲۴٫۵۹۹	۷۷٫۳۷۲	۰٫۹۹۹۵
		انحراف معیار	۰٫۰۰۳۴	۰٫۰۰۰۶	۰٫۰۰۰۶	۱٫۸۹۳۷	۱٫۸۹۳۷	۰٫۰۰۱۶
	۰٫۳	میانگین	۰٫۹۱۴۱	۰٫۹۲۰۶	۰٫۹۸۲۰	۷۱٫۶۵۷	۱۲۴٫۴۳	۰٫۹۵۱۵
		انحراف معیار	۰٫۰۱۶۴	۰٫۰۰۳۶	۰٫۰۰۳۸	۹٫۰۹۹۹	۹٫۰۹۹۹	۰٫۰۰۹۱
	۰٫۵	میانگین	۰٫۷۳۵۱	۰٫۸۶۹۳	۰٫۹۲۷۲	۱۷۰٫۹۲	۲۲۳٫۶۹	۰٫۸۳۹۵
		انحراف معیار	۰٫۰۳۱۳	۰٫۰۱۱۴	۰٫۰۱۲۱	۱۷٫۳۵۴	۱۷٫۳۵۴	۰٫۰۲۰۵

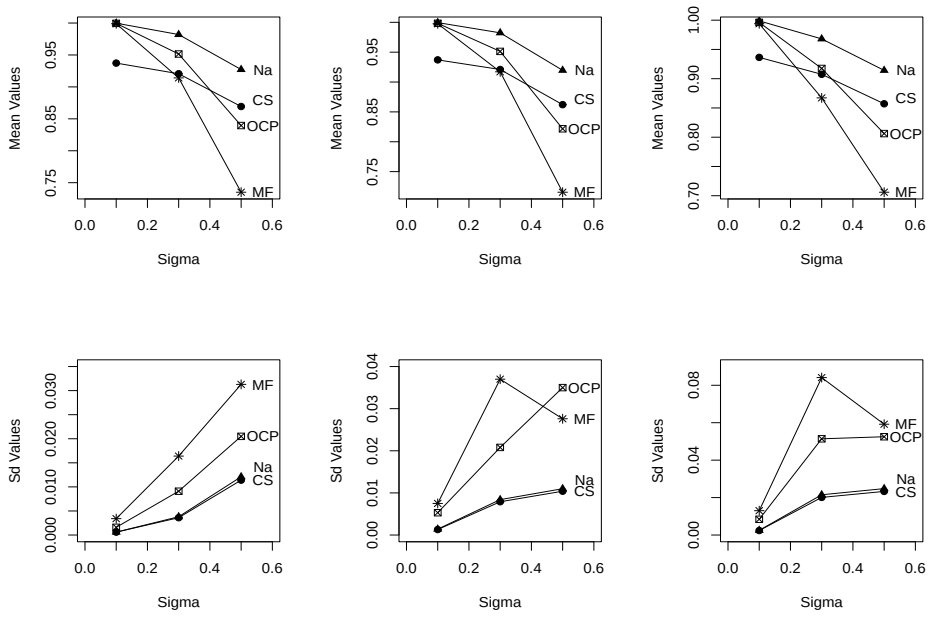
نتایج حاصل از رده‌بندی در قالب معیارهای ضریب تعیین-نمای[1] مک‌فادن، کاکس‌اسنیل و ناگلکرک و همچنین معیارهای AIC، BIC و درصد پیش‌بینی صحیح کلی (OCP) به‌دست آمد. میانگین و انحراف معیار این نتایج در جدول ۱ برای مقادیر مختلف از σ محاسبه شده است. همانگونه که قابل مشاهده است با افزایش میزان σ میانگین در تمامی ضرایب تعیین‌نما کاهش می‌یابد در صورتی که معیارهای AIC و BIC افزایش چشم‌گیری دارند. همچنین مقادیر انحراف معیار نیز افزایش می‌یابد. خلاصه‌ای از مقادیر جدول ۱ در قالب نمودارها در شکل ۴ به‌نمایش گذاشته شده است. اصلی‌ترین معیار برای بررسی عملکرد مدل را می‌توان معیار OCP در نظر گرفت که میزان رده‌بندی درست را نشان می‌دهد. همانگونه که انتظار می‌رود با افزایش σ مقدار OCP کاهش می‌یابد. با افزایش تعداد پیکربندی‌ها در هر رده میزان دقت در رده‌بندی بر اساس معیار OCP بهبود می‌یابد.

۴.۲ کاربرد

دو مجموعه داده واقعی مورد استفاده در این مقاله از بسته shapes در نرم‌افزار R گرفته شده‌اند:

اسکیزوفرنی (Schi): برای ثبت این داده‌ها، ابتدا تصویربرداری رزونانس مغناطیسی (MRI) در نزدیکی میدساژیتال[2] از ۱۴ بیمار اسکیزوفرنی و ۱۴ فرد کنترل انجام گرفت. سپس برش‌های دو بعدی از تصاویر در نظر گرفته و ۱۳ نقطه

[1] Pseudo-R-Square
[2] Midsagittal



شکل ۴. از راست به چپ: نمودارهای میانگین (بالا) و انحراف معیار (پایین) ضرایب تعیین‌نمای مک‌فادن، کاکس‌اسنیل، ناگلکرک، و معیار OCP برای ۱۰، ۳۰ و ۵۰ پیکربندی براساس ۱۰۰ بار تکرار شبیه‌سازی.

شاخص آن‌ها ثبت شد. این داده ها در دو بعد هستند(بوکشتین ، ۱۹۹۶).

میمون‌ها: داده‌ها مربوط به پیکربندی جمجمه ۱۶۷ میمون بالغ، شامل ۳۰ گوریل ماده ،(gorf) ۲۹ گوریل نر (gorm)، ۲۶ شامپانزه ماده ،(panf) ۲۸ شامپانزه نر (panm)، ۲۴ اورانگوتان ماده (pongof) و ۳۰ اورانگوتان نر (pongom) است (اهیگینز ، ۱۹۸۹؛ اهیگینز و درایدن ، ۱۹۹۳).

برای رده‌بندی این دو مجموعه داده، ابتدا پارامتر پهنای باند h را با استفاده از معیار CV برآورد می‌کنیم. با استفاده از $\hat{h}$، مدل MLR را برازش داده و پیکربندی‌ها را رده‌بندی می‌کنیم. نتایج در جدول ۲ نشان داده شده است. توجه داشته باشید که برای محاسبه معیارهای گزارش شده در جدول ۲ هر مشاهده را کنار گذاشته و سپس رده آن مشاهده را با استفاده ازسایر داده‌ها پیش‌بینی کردیم. برای ۲۸ داده‌ی اسکیزوفرنی، رده مربوط به ۱۲ فرد در رده کنترل و ۱۱ فرد در رده بیمار به‌درستی پیش‌بینی شد. همچنین رده ۲ فرد سالم و ۳ فرد بیمار اشتباه پیش‌بینی شد. برای مجموعه داده میمون‌ها نیز ماتریس درهم‌ریختگی[1] آن‌ها ایجاد شد. این ماتریس در جدول ۳ قابل مشاهده است. مقادیر داخل جدول تعداد پیش‌بینی‌ها بر اساس رده واقعی هر میمون و رده پیش‌بینی شده را نشان می‌دهد.

[1] Confusion matrix

جدول ۲. معیارهای ضریب تعیین‌نما مک‌فادن، کاکس‌اسنیل و ناگلکرک وخلاصه‌ای از برازش رگرسیون لوژستیک شامل AIC، BIC و OCP برای داده‌های اسکیزوفرنی و میمون‌ها.

مجموعه داده	ضریب تعیین‌نما			خلاصه مدل		
	مک‌فادن	کاکس‌اسنیل	ناگلکرک	AIC	BIC	OCP	$\hat{h}$
اسکیزوفرنی	۰٫۲۰	۰٫۲۴	۰٫۳۲	۴۰٫۴۹	۳۱٫۱۶	۸۲	۲۵٫۷
میمون‌ها	۰٫۸۲	۰٫۹۵	۰٫۹۷	۳۵٫۱۶	۲۱۱٫۴	۸۴	۰٫۰۸

جدول ۳. ماتریس درهم‌ریختگی گروه‌های میمون‌ها با رگرسیون لوژستیک چندجمله‌ای نیمه‌پارامتری.

مشاهده	پیش‌بینی						رده‌بندی درست (درصد)
	gorf	gorm	panf	panm	pongof	pongom	
gorf	۳۰	۰	۰	۰	۰	۰	۱۰۰
gorm	۰	۲۹	۰	۰	۰	۰	۱۰۰
panf	۰	۰	۱۶	۱۰	۰	۰	۶۱٫۵
panm	۰	۰	۱۰	۱۷	۲	۰	۶۳٫۰
pongof	۰	۰	۰	۱	۲۰	۲	۹۰٫۹
pongom	۰	۰	۰	۰	۲	۲۸	۹۳٫۳
مجموع (درصد)	۱۸	۱۷٫۴	۱۵٫۶	۱۶٫۲	۱۳٫۲	۱۸	۸۴

جدول ۴. معیار RMS برای هر دو گروه جمجمه گوریل‌های نر و ماده قبل و بعد از رده‌بندی طبق روش‌های (الف) و (ب) و ارائه شده در این مقاله (ج).

		قبل از رده‌بندی	بعد از رده‌بندی (الف)	بعد از رده‌بندی (ب)	بعد از رده‌بندی (ج)
RMS	ماده	۰٫۰۴۳۷	۰٫۰۴۶۹	۰٫۰۴۳۲	۰٫۰۴۳۷
	نر	۰٫۰۴۹۹	۰٫۰۴۶۵	۰٫۰۴۷۸	۰٫۰۴۹۹
میانگین دو RMS		۰٫۰۴۶۸	۰٫۰۴۶۷	۰٫۰۴۶۸	۰٫۰۴۶۸
فاصله بین میانگین‌ها		۰٫۰۵۸۶	۰٫۰۵۹۰	۰٫۰۵۹۴	۰٫۰۵۸۶

به منظور مقایسه روش معرفی شده در این مقاله با روش‌های دیگر رده‌بندی پیکربندی‌ها، ما روش خود را با روش‌های ارائه شده در الف-نبیل و گل‌علی‌زاده (۲۰۱۶) و ب- مقیم بیگی و نودهی (۲۰۲۲) بر اساس اندازه‌گیری RMS و داده‌های جمجمه گوریل (gorf و gorm) مقایسه می‌کنیم. این نتایج در جدول ۴ نشان داده شده است. براساس معیار RMS اگر مقدار RMS در هر رده، قبل و بعد از رده‌بندی نزدیک باشد، رده‌بندی به‌طور موثر انجام شده است. در واقع این مقدار نشان می‌دهد که میزان همگنی قبل و بعد از رده‌بندی به چه اندازه به یکدیگر نزدیک بوده است. همان‌طور که در این جدول مشاهده می شود، مقادیر RMS در مدل ما قبل و بعد از رده‌بندی بسیار نزدیک به هم هستند که نشان می‌دهد مدل ارائه شده در این مقاله (ج) از روش‌های (الف) و (ب) عملکرد بهتری در رده‌بندی این مجموعه از داده‌ها دارد. در واقع یکسان بودن مقادیر جدول قبل و بعد از رده‌بندی با استفاده از روش ارائه شده در این مقاله نشان دهنده این است که رده‌بندی به‌طور کامل و با OCP برابر ۱۰۰ درصد انجام گرفته است.

بحث و نتیجه‌گیری

در این مقاله رده‌بندی پیکربندی‌ها با استفاده از یک مدل نیمه پارامتری انجام گرفت. بخش ناپارامتری مدل با به حداقل رساندن معیار توان-واگرایی موضعی، که در آن تابع وزن با استفاده از توزیع آماری فضای شکل به دست آمده بود، معرفی شد. پارامتر هموارسازی نیز با استفاده از روش اعتبارسنجی متقابل حداقل مربعات برآورد شد. با توجه به این‌که برآورد ناپارامتری یک برآورد اریب است برای رده‌بندی پیکربندی‌ها یک مدل MLR معرفی شد. در این مدل، برآورد پارامترهای MLR با استفاده از روش شبکه عصبی انجام گرفت. سپس پیکربندی‌های شبیه‌سازی شده با استفاده از مدل معرفی شده رده‌بندی شدند. همچنین دو مجموعه داده واقعی نیز با استفاده از مدل رده‌بندی شدند که نشان از عملکرد مناسب مدل داشت. بر اساس تحلیل ارائه شده در زیربخش ۴.۲، روش معرفی شده در این مقاله در مقایسه با روش‌های ارائه شده توسط مقیم بیگی و نودهی (۲۰۲۲) و نبیل و گل‌علی‌زاده (۲۰۱۶) از عملکرد مناسب‌تری برخوردار است. اما از نقطه نظر محاسباتی باید به این نکته اشاره کرد که روش نیمه پارامتری ارائه شده در این مقاله زمان‌بر است که هدف تحقیق آتی نویسندگان این مقاله رویارویی با این نقیصه است.

تقدیر و تشکر

نویسنده مقاله از رهنمودها و نظرات ارزنده داوران، سردبیر و ویراستار محترم مجله علوم آماری که سبب ارتقای کیفی مقاله شد، کمال تشکر و قدردانی را دارد.

مراجع

مقیم بیگی، م. و گل‌علی‌زاده م. (۱۳۹۸)، مدل‌بندی رگرسیونی شکل از طریق مثلثی کردن، مجله علوم آماری، ۱۳(۱)، ۱۸۵-۱۹۶.

فتوحی، ح. و گل‌علی‌زاده، م. (۱۳۹۱)، بهبود عملکرد تحلیل ژئودزیک اصلی در تحلیل آمار شکل، مجله علوم آماری، ۲(۶)، ۲۱۹-۲۳۶.

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle. In *B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory.*

Bookstein, F. L. (1996), Biometrics, Biomathematics and the Morphometric Synthesis, *Bulletin of Mathematical Biology*, **58**, 313–365.

Bookstein, F. L. (1986), Size and Shape Spaces for Landmark Data in Two Dimensions (With Discussion), *Statist. Sci.*, **1**, 181–242.

Cressie, N. and Read, T. R. (1984), Multinomial Goodness-of-Fit Tests, *Journal of the Royal Statistical Society: Series B (Methodological)*, **46**, 440–464.

Cox, D. R. and Snell, E. J. (1989), *Analysis of Binary Data*, Chapman and Hall/CRC, Boca Raton.

Debavelaere, V. Durrleman, S. and Allassonnière, S. (2020), Learning the Clustering of Longitudinal Shape Data Sets into a Mixture of Independent or Branching Trajectories, *International Journal of Computer Vision*, **128**, 2794–2809.

Dryden, I. L. (2012), *Shapes package*, R Foundation for Statistical Computing, Vienna, Austria, Contributed package.

Dryden, I. L. Hirst, J. D. and Melville, J. L. (2007), Statistical Analysis of Unlabeled Point Sets: Comparing Molecules in Chemoinformatics, *Biometrics*, **63**, 237–251.

Dryden, I. L. and Mardia, K. V. (2016), *Statistical Shape Analysis: With Applications in R*, John Wiley and Sons, Chichester.

Duin R. P. W. (1976), On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions, *IEEE Transactions on Computers C-25*, 1175–1179.

Kendall, D. G. (1977), The Diffusions 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.

Lin, K. C. and Chen, Y. J. (2008), Assessing Ordinal Logistic Regression Models via Nonparametric Smoothing, *Communications in Statistics-Theory and Methods*, **37**, 917–930.

Moghimbeygi, M. and Nodehi, A. (2022), Multinomial Principal Component Logistic Regression on Shape Data, *Journal of Classification*, 1–22.

Nabil, M. and Golalizadeh, M. (2016), On Clustering Shape Data. *Journal of Statistical Computation and Simulation*, **86**, 2995–3008.

O'Higgins, P. (1989), *A Morphometric Study of Cranial Shape in the Hominoidea*, PhD thesis, University of Leeds .

O'Higgins, P. and Dryden, I. L. (1993), Sexual Dimorphism in Hominoids: Further Studies of Craniofacial Shape Differences in Pan, Gorilla, Pongo, *Journal of Human Evolution*, **24**, 183–205.

Okumura, H. and Naito, K. (2006), Nonparametric Kernel Regression for Multinomial Data, *Journal of multivariate analysis*, **97**, 2009–2022.

Rifada, M. Chamidah, N. and Ratnasari, V. (2021), Estimation of Nonparametric Ordinal Logistic Regression Model Using Local Maximum Likelihood Estimation, *Commun. Math. Biol. Neurosci.,* 2021, Article-ID.

Schwarz, G. (1978), Estimating the Dimension of a Model, *The annals of statistics*, **6**, 461–464.

Simó, A. Victoria Ibáñez, M. Epifanio, I. and Gimeno, V. (2020), Generalized Partially Linear Models on Riemannian Manifolds, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **69**, 641–661.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S. 4th ed.* Springer.

Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer Science & Business Media.

Yankov, D. and Keogh, E. (2006), Manifold Clustering of Shapes, *In Proc. IEEE Int. Conf. Data Mining (ICDM).*