# تُنُک‌سازی مؤلفه‌های اصلی در حضور نقاط دورافتاده

امیر رزاقی،[۱] موسی گلعلی‌زاده،[۲]

*تاریخ دریافت: ۱۳۹۷/۱۱/۲۸*
*تاریخ پذیرش: ۱۳۹۸/۷/۲۰*

چکیده:
یکی از معروف‌ترین رویکردهای اکتشافی برای کاهش بُعد و توصیف ساده‌تر منابع اصلی تغییرات، تحلیل مؤلفه‌های اصلی است. با وجود مزایای جالب توجه این روش، به‌کارگیری آن در برخی از مواقع مشکلاتی را به همراه دارد. حضور نقاط دورافتاده در مجموعه‌داده‌ها، تأثیرهای مخربی بر نتایج این رویکرد دارد که به نظر می‌رسد گونه‌ای از مؤلفه‌های اصلی که اُستوار باشند برای اخذ نتایج معتبر، سودمند است. به علاوه، وجود بارهای میانی در برخی از ترکیبات خطی، تفسیر مؤلفه‌ها را دشوار می‌سازد که در این حالت می‌توان گونه‌ای از تُنُک‌سازی مؤلفه‌ها را در نظر گرفت. در این مقاله، برای حصول هم‌زمان مؤلفه‌های اصلی اُستوار و تُنُک، رویکرد ترکیبی کارآمدی ارائه و سپس به‌منظور ارزیابی و مقایسه آن با رویکردهای مطرح شده از شبیه‌سازی آماری بهره گرفته می‌شود. در نهایت، ابزارهای مورد اشاره در تحلیل مثال واقعی مرتبط با مجموعه‌داده‌های جرم و جنایت در آمریکا مورد استفاده قرار می‌گیرد.

**واژه‌های کلیدی:** تحلیل مؤلفه‌های اصلی، نقاط دورافتاده، تعبیرپذیری مؤلفه‌ها، تحلیل مؤلفه‌های اصلی اُستوار و تُنُک، داده‌های جرم و جنایت.

## ۱ مقدمه

یکی از مسائل مهم در تحلیل داده‌هایی با مجموعه زیادی از متغیر، خلاصه نمودن متغیرها به‌منظور اخذ نتایج مفید از داده‌های اوّلیه است. از این رو، رویکردی چندمتغیره با هدف کاهش بُعد و توصیف تغییرات موجود در مجموعه‌داده‌ها روی کار آمد که به رویکرد تحلیل مؤلفه‌های اصلی[۳](PCA) معروف است (جولیفه، ۲۰۰۲).

علی‌رغم مزیت‌های فراوان ابزار، PCA این رویکرد در برخی از موارد، نتایج معتبری را برای محقق به همراه ندارد (اوریت و هاترن، ۲۰۱۱). بنا به هابرت و دیگران (۲۰۱۶)، دو مشکل اساسی در به‌کارگیری رویکرد، PCA وجود نقاط دورافتاده[۴] و عدم تعبیرپذیری[۵] مناسب مؤلفه‌های اصلی است. وجود نقاط دورافتاده در داده‌ها بر بردار میانگین اثر مخربی گذاشته و منجر به بی‌اعتباری نتایج حاصل می‌شود. به نظر می‌رسد اتخاذ روشی اُستوار که به طریق مناسبی با رویکرد PCA تلفیق شده باشد، راه حل معقولی باشد. مشکل دیگر چنین رویکردی، پیچیدگی تفسیر مؤلفه‌های اصلی به‌دست‌آمده است. این امر ناشی از آن است که اکثریت بارها[۶] از لحاظ قدر مطلق، اعدادی نه خیلی کوچک و نه خیلی بزرگ هستند. به نظر می‌رسد در این حالت، گونه‌ای از تُنُک‌سازی مؤلفه‌ها که باعث سوق تعدادی از ضرایب

[۱] دانشجوی کارشناسی ارشد آمار، دانشگاه تربیت مدرس، تهران، ایران
[۲] هیأت علمی گروه آمار، دانشگاه تربیت مدرس، تهران، ایران

[۳] Principal components analysis
[۴] Outliers
[۵] Interpretability
[۶] Loadings

متغیرها به سمت صفر شود، مفید فایده است. اکنون اگر حالت ترکیبی وجود نقاط دورافتاده و دشواری تعابیر مؤلفه‌ها مد نظر قرار گیرد، آن‌گاه انتظار می‌رود تلفیق دو راه حل اشاره شده، رویکرد مناسبی در برخورد با مشکل اخیر باشد. بررسی منابع علمی نشان می‌دهد که اوّلین تحقیق در این‌گونه مسائل، توسط کروکس و دیگران (۲۰۱۳) صورت گرفت که حاصل تلاش آنها ارائه رویکرد تحلیل مؤلفه‌های اصلی اُستوار تُنُک[۷](SRPCA) بود. در ادامه نیز، هابرت و دیگران (۲۰۱۶)، رویکرد تحلیل مؤلفه‌های اصلی تُنُک اُستوار[۸](ROSPCA) را پیشنهاد دادند. مقاله حاضر، مروری بر فعالیت‌های ذکر شده و سپس کاربست رویکرد ROSPCA در شبیه‌سازی و مثال واقعی است.

برای ارائه مفاهیم مرتبط با موضوعات مطرح شده در بخش مقدمه، ادامه مقاله به بخش‌های زیر تقسیم شده است. در بخش دوم، مقدمه‌ای از تحلیل مؤلفه‌های اصلی اُستوار ارائه و سپس روش‌های اُستوارسازی مرور می‌شود. جزئیاتی راجع به روش SRPCA و تشریحی از رویکرد ترکیبی ROSPCA به‌منظور اُستوارسازی و تُنُک‌سازی مؤلفه‌های اصلی به‌طور هم‌زمان در بخش سوم ارائه می‌شود. در بخش چهارم، برای ارزیابی رویکرد ROSPCA با رویکردهای قبلی، مطالعه‌ای شبیه‌سازی صورت می‌گیرد و کاربست آن در تحلیل یک مثال واقعی مورد بررسی قرار می‌گیرد. نتیجه‌گیری کلّی از مطالب مطرح شده، پایان‌بخش مقاله حاضر است.

# ۲ اُستوارسازی مؤلفه‌های اصلی

حساسیت ماتریس کواریانس به نقاط دورافتاده و در نتیجه تأثیر مخرب آنها بر ماتریس بار، یکی از مشکلات رویکرد PCA[ کلاسیک است. در واقع، حضور چنین نقاطی سبب می‌شود که مؤلفه‌ها به سمت نقاط دورافتاده کشیده شده و نتوانند تغییرات مشاهدات منظم را به‌درستی منعکس کنند (هابرت و دیگران، ۲۰۰۵). از این رو، روش‌های اُستوار مختلفی پیشنهاد شده‌اند که در ادامه تشریح می‌شوند.

## ۱.۲ تحلیل مؤلفه‌های اصلی اُستوار با استفاده از ماتریس کواریانس اُستوار

یکی از مهم‌ترین رویکردهای اُستوارسازی مؤلفه‌های اصلی، استفاده از ماتریس کواریانس اُستوار است. ساخت این ماتریس بر مبنای برآوردگرهایی است که مکان[۹] و پراکنش اُستواری را تولید کنند. یکی از نخستین برآوردگرهای هم‌وردای آفینی برای به دست آوردن این دو معیار در تحلیل چندمتغیره، برآوردگر مینیمم دترمینان کواریانس[۱۰](MCD) است (روسیوف، ۱۹۸۵). طریقه به دست آوردن برآوردگر MCD بر مبنای یافتن زیرنمونه‌هایی به اندازهٔ $h$ از مجموعه مشاهدات به اندازهٔ $n$ است که ماتریس کواریانس معمول این $h$ مشاهده، کمترین دترمینان ممکن را داشته باشد. از دیدگاه ریاضی، اندازه زیرنمونه‌ها ($h$) از رابطهٔ

$$h = h_\alpha = h(\alpha, n, p) = [۲n_۲ - n + ۲\alpha(n - n_۲)], \quad (۱)$$

به دست می‌آید که در آن $n_۲ = [\frac{n+p+۱}{۲}]$، $n$ تعداد کل مشاهدات، $p$ تعداد متغیرهای موجود و $\alpha$ پارامتر عددی کنترل‌کنندهٔ اندازه زیرنمونه‌ها برای مینیمم شدن دترمینان است که مقادیری بین ۰٫۵ تا ۱ را اختیار می‌کند. شایان ذکر است که در تساوی (۱) و همچنین تعریف $n_۲$، نماد [] نشان‌دهندهٔ جزء صحیح است. اگر چه تعیین مقدار $\alpha$ به سلیقه محقق و طبیعت مسئله بر می‌گردد ولی روسیوف و دریزن (۱۹۹۹) نشان دادند که بیشترین اُستواری برای حالت $\alpha = ۰٫۵$ اتفاق می‌افتد. بنا بر این زمانی که $\alpha = ۰٫۵$، با فراخوانی رابطهٔ (۱)، اندازه زیرنمونه‌ها از رابطهٔ $h_{۰٫۵} = [\frac{n+p+۱}{۲}]$ به دست می‌آید. در نهایت، برآورد مکان و پراکنش چندمتغیره برآوردگر MCD با پارامتر $h$، به‌صورت زیر حاصل می‌شوند:

[۷] Sparse robust PCA

[۸] Robust sparse PCA

[۹] Location

[۱۰] Minimum covariance determinant

الف) کمّیت $\hat{\mu}_0$، میانگین $h$ مشاهده‌ای است که دترمینان ماتریس کواریانس آن مینیمم خواهد بود.
ب) $\hat{\Sigma}_0$، ماتریس کواریانس مربوط به $h$ مشاهده با احتساب ضرب در عامل سازگاری[۱۱] ($c_0$) است. بنا به کروکس و دیگران (۱۹۹۹) در توزیع نرمال، $c_0 = \frac{\alpha}{F_{\chi^2_{p+2}}(\chi^2_{p,\alpha})}$ که در آن $\alpha = \lim_{n\to\infty} \frac{h(n)}{n}$.

شده ماکسیمم باشد. لذا، برای حصول مؤلفه‌های اصلی اُستوار در روش‌های PP از شاخص تصویری اُستوار استفاده می‌شود. لازم به اشاره است که برآوردگرهای میانه انحراف مطلق[۱۴] (MAD) و $Q_n$ از جمله برآوردگرهای مقیاس اُستوار هستند که می‌توان از مجذور آنها به‌عنوان شاخص‌های تصویر استفاده کرد.

## ۲.۲ اُستوارسازی مؤلفه‌های اصلی بر اساس جستجوی تصویر

علی‌رغم کارایی مناسب رویکرد PCA بر مبنای ماتریس کواریانس اُستوار، این روش دارای اشکالاتی است که هابرت و دیگران (۲۰۰۵) به آنها اشاره کردند. مشکل نخست، به دست آوردن ماتریس کواریانس اُستوار در داده‌های با بُعد بالا به دلیل محاسبات پیچیده آن است. مشکل دیگر این است که در برخی از مسائل، اندازه نمونه از تعداد متغیرها کمتر است ($n < p$). در این حالت، محاسبه برآوردگر MCD ممکن نیست چرا که این برآوردگر در صورتی محاسبه می‌شود که تعداد متغیرهای موجود از $h$ مشاهده، که در واقع اندازه زیرنمونه در رابطهٔ (۱) است، کوچک‌تر باشد یعنی $p < h$. واضح است که در غیر این صورت ($p > h$)، دترمینان ماتریس کواریانس هر $h$ زیرنمونه به دلیل رتبه کامل نبودن برابر صفر خواهد بود.
برای رفع مشکلات مذکور، می‌توان از ایده هابر (۱۹۸۵) که رویکردی تحت عنوان جستجوی تصویر[۱۲] (PP) پیشنهاد داد، استفاده کرد. هدف اصلی روش‌های، PP ارائه تصاویری با بُعد کم از مجموعه نقاطی با بُعد بالا به‌وسیله ماکسیمم‌سازی عددی تابع هدفی مشخص است که به این تابع هدف، شاخص تصویر[۱۳] می‌گویند. در واقع، داده‌ها بر روی زیرفضایی با بُعد پایین تصویر می‌شوند به‌گونه‌ای که مقدار اُستواری پراکنش داده‌های تصویر

## ۳.۲ تحلیل مؤلفه‌های اصلی اُستوار تلفیقی

به‌منظور اُستوارسازی مؤلفه‌های اصلی که از ویژگی‌های مفید دو روش نامبرده برخوردار باشد، رویکرد تحلیل مؤلفه‌های اصلی اُستوار (ROBPCA) توسط هابرت و دیگران (۲۰۰۵) معرفی شد. این رویکرد از روش‌های PP برای کاهش بُعد اوّلیه متغیرها و از برخی ویژگی‌های برآوردگر MCD برای فضای داده با بُعد کمتر بهره می‌برد. در این رویکرد، با فرض این که $n$ تعداد مشاهدات و $p$ تعداد متغیرهای موجود در ماتریس داده $X$ باشد، ابتدا فضای کلّی داده‌ها با استفاده از تجزیه ویژه مقدار[۱۵] (SVD) ماتریس داده به زیرفضای آفینی پدید آمده[۱۶] توسط $n$ مشاهده کاهش می‌یابد و مجموعه‌داده‌ها در زیرفضایی با بُعد حد اکثر $n-1$ قرار می‌گیرند. سپس، ماتریس کواریانس اوّلیه $\Sigma_h$ برای تعیین $k$ مؤلفه اصلی و در نتیجه زیرفضای $k$ بُعدی ساخته می‌شود. سرانجام، مجموعه نقاط داده بر روی زیرفضای مذکور که مکان و پراکنش آن به‌صورت اُستوار برآورد شده است، تصویر شده و $k$ ویژه بردار $l_1, l_2, ..., l_k$ به دست می‌آید. اکنون، ویژه بردارهای مربوط به این ویژه مقادیر، $k$ مؤلفه اصلی اُستوار هستند.
خواننده علاقه‌مند به مطالعه جزئیات کامل و دقیق رویکرد ROBPCA و گام‌های اجرایی آن می‌تواند به هابرت و دیگران (۲۰۰۵) مراجعه کند.

[۱۱] Consistency factor
[۱۲] Projection pursuit
[۱۳] Projection index
[۱۴] Median absolute deviation
[۱۵] Singular-value decomposition
[۱۶] Affine subspace spanned

# ۳ تحلیل مؤلفه‌های اصلی اُستوار و تُنُک

یکی از جنبه‌های مشکل مجموعه توابع خطی در رویکردهای چندمتغیره مانند، PCA تعبیر و تفسیر آنها است. لذا، برای حل این مشکل و به دست آوردن مؤلفه‌های تعبیرپذیر، رویکردهای تُنُک متفاوتی پیشنهاد شده‌اند که در این بخش به توصیف مختصری از رویکرد تحلیل مؤلفه اصلی ساده‌سازی شده بر مبنای روش کمترین انقباض مطلق و عملگر انتخاب[۱۷](SCoTLASS) پرداخته می‌شود (جولیفه و دیگران، ۲۰۰۳). سپس، برای حصول مؤلفه‌هایی که ویژگی اُستواری و تُنُکی را به‌صورت توأم به همراه دارند، رویکرد ROSPCA تشریح می‌شود.

## ۱.۳ تُنُک‌سازی مؤلفه‌های اصلی مبتنی بر LASSO

به سبب وجود تعداد زیاد متغیرهای پیشگو[۱۸] در معادلات رگرسیونی چندگانه و مشکلات ناشی از آن، تیبشیرانی (۱۹۹۶) برای رفع مشکل تفسیر چنین معادلاتی، مطالعه‌های گسترده‌ای انجام داد و توانست ابزار مفیدی پیشنهاد کند. این رویکرد ارائه شده که توافقی میان انتخاب متغیر[۱۹] و برآوردگرهای انقباضی[۲۰] است، تحت عنوان "کمترین انقباض مطلق و عملگر انتخاب"[۲۱](LASSO) معرفی شد.

جولیفه و دیگران (۲۰۰۳) نیز به‌منظور تسهیل تفاسیر مؤلفه‌های اصلی، رویکردی بر مبنای روشی رگرسیون‌گونه پیشنهاد کردند. در واقع، آنها بر این عقیده بودند که می‌توان رویکرد PCA را به‌صورت یک مسئله بهینه‌سازی ساختاربندی نمود، آن‌گاه با استفاده از ایجاد قید LASSO بر روی ضرایب آن، ماتریس بار تُنُک و در نتیجه مؤلفه‌های اصلی قابل تعبیری ایجاد کرد. شایان ذکر است که رویکرد SCoTLASS بر اساس ماکسیمم‌سازی متوالی واریانس مؤلفه‌های اصلی تحت قید اضافی $\sum_{j=1}^{p}|a_{kj}| \le t$ با توجه به پارامتر تنظیم‌کنندهٔ $t$ صورت می‌گیرد که $a_{kj}$ نشان‌دهندهٔ $j$اُمین عنصر از $k$اُمین بردار $a$ است.

## ۲.۳ تعبیرپذیری مؤلفه‌های اصلی اُستوار

وجود نقاط دورافتاده و پیچیده بودن تفاسیر مؤلفه‌های اصلی به‌طور هم‌زمان سبب شد که محققان درصدد ارائه رویکردهای مناسب برای رفع چنین مشکلی باشند. در ابتدا، کروکس و دیگران (۲۰۱۳) با استفاده از ادغام قواعد تُنُکی به ساختار روش‌های PP توانستند رویکرد ترکیبی SRPCA را ارائه کنند. اما از آن‌جا که شناسایی نقاط دورافتاده در مسائل چندمتغیره از اهمیت بالایی برخوردار است، محققان رویکرد جدیدی ارائه کردند. در واقع بر خلاف دیدگاه، SRPCA هابرت و دیگران (۲۰۱۶) رویکردی کارآمد و مفید تحت عنوان ROSPCA پیشنهاد دادند که از ویژگی‌های رویکرد ROBPCA کمک می‌گیرد.

به‌طور کلّی، ایدهٔ اصلی رویکرد ROSPCA مبتنی بر اجرای جداگانه کشف نقاط دورافتاده و تُنُک‌سازی مؤلفه‌ها است اما نتیجه نهایی کار به طریقی، هدف مشترک مقابله با نقاط دورافتاده و تعبیرپذیری مناسب مؤلفه‌ها را تأمین می‌کند. می‌توان گفت که از دیدگاه محاسباتی، این رویکرد در ابتدا همانند روش اُستوارسازی ROBPCA به شناسایی نقاط دورافتاده در مجموعه‌داده‌ها پرداخته و سپس با پیروی از رویکرد SCoTLASS مبتنی بر الگوریتم تور، تُنُک‌سازی مؤلفه‌ها را به مرحله اجرا در می‌آورد. با اجرای دو مرحله مورد اشاره، شخص با ماتریس کواریانسی روبرو می‌شود که ویژه بردارهای آن به‌درستی تغییرات بین داده‌های مورد مطالعه را توصیف می‌کنند. به بیانی دیگر، ویژه بردارهای حاصل در مقابل نقاط دورافتاده اُستوار بوده و در عین حال تعبیر مناسبی برای مؤلفه‌های اصلی ارائه

[۱۷] Simplified component technique-least absolute shrinkage and operator

[۱۸] Predictor variables

[۱۹] Variable selection

[۲۰] Shrinkage estimators

[۲۱] Least absolute shrinkage and selection operator

می‌کنند. جزئیات بیشتر و کامل‌تری از گام‌های موجود برای اجرای رویکرد ROSPCA در هابرت و دیگران (۲۰۱۶) موجود است.

قابل ذکر است که رویکرد ROSPCA شامل دو ابر پارامتر[۲۲] به‌صورت $\alpha$ و $\lambda$ است که این مقادیر به‌ترتیب درجه اُستواری و میزان پارامتر تُنُکی داده‌ها را تعیین می‌کنند. گفتنی است که مقدار $\alpha$، کران پایینی بر روی تعدادی از مشاهدات منظم ایجاد می‌کند و این موضوع بدان معناست که با تعیین مقدار آن، حد اکثر $(۱-\alpha)۱۰۰\%$ از $n$ مشاهده می‌توانند به‌عنوان نقاط دورافُتاده در نظر گرفته شوند. به علاوه، بنا به هابرت و دیگران (۲۰۱۶)، مقدار $\lambda$ بر اساس مینیمم‌سازی گونه‌ای از معیار اطلاع بیزی که مبتنی بر مجموع توان‌های دوم مانده‌ها است، انتخاب می‌شود. این معیار به‌صورت

$$\mathrm{BIC}(\lambda) = \ln(\frac{۱}{h_۱ p}\sum_{i=۱}^{h_۱} OD_{(i)}^{۲}(\lambda)) + \mathrm{df}(\lambda)\frac{\ln(h_۱ p)}{h_۱ p}, \quad (۲)$$

تعریف می‌شود که در آن $h_۱$ اندازه مجموعه $H_۱$ ظاهر شده در الگوریتم، ROSPCA $p$ تعداد متغیرهای موجود، $OD_{(i)}(\lambda)$ کوچک‌ترین فاصله متعامد $i$اُم مدل در حضور پارامتر تُنُکی $(\lambda)$ و $\mathrm{df}(\lambda)$ تعداد بارهای غیر صفر زمانی که از $\lambda$ به‌عنوان پارامتر تُنُکی استفاده شده است، هستند. با توجه به رابطهٔ (۲) می‌توان دریافت که بخش نخست آن، میزان دقت و کیفیت برازش مدل را اندازه‌گیری می‌کند در حالی که بخش دوم، نشانگر تاوانی برای پیچیدگی مدل است و سبب تعاملی میان درستی و تُنُکی مدل می‌شود. در عمل، مقدار $\lambda$ به‌وسیلهٔ مینیمم‌سازی معیار مذکور روی بازه $[۰, \lambda_{\max}]$ انتخاب می‌شود که $\lambda_{\max}$ معرف تُنُکی کامل (دقیقاً یک بار غیر صفر در هر مؤلفه) است. برای تحقق این امر، مقدار بهینه $\lambda$ برای رویکرد مربوط به‌وسیلهٔ بررسی تور هم‌فاصله[۲۳] مقادیر $\lambda$ روی این بازه انتخاب می‌شود. به‌عبارت دقیق‌تر می‌توان گفت که به‌ازای هر مقدار $\lambda$، رویکرد مورد نظر با احتساب مقدار پارامتر تُنُکی مفروض بر روی مجموعه‌داده مورد مطالعه اعمال و سپس معیار اطلاع[۲۴](IC) محاسبه می‌شود. سرانجام، مقدار بهینه $\lambda$ مقداری است که به‌ازای آن، کمترین مقدار IC رابطهٔ (۲) حاصل شود.

# ۴ شبیه‌سازی آماری و تحلیل مثال واقعی

در این بخش، ابتدا با استفاده از انجام شبیه‌سازی به ارزیابی عملکرد رویکرد ROSPCA پرداخته و بر اساس معیارهایی از قبیل زاویه بین زیرفضاها[۲۵] و معیار میانگین کل صفرها[۲۶] که مناسب ارزیابی روش‌های مرتبط با PCA هستند، رویکردهای مطرح شده با یکدیگر مقایسه می‌شوند. سپس در ادامه، نحوه کاربست رویکرد ROSPCA در یک مثال واقعی تشریح می‌شود. گفتنی است که در این مقاله، انجام شبیه‌سازی و محاسبات آن با بسته آماری rospca (رینکینز، ۲۰۱۸) در نرم‌افزار R قابل انجام است.

## ۱.۴ بررسی شبیه‌سازی

به‌منظور ارزیابی رویکرد ROSPCA و مقایسه آن با رویکردهای، CPCA، SCoTLASS ROBPCA و SRPCA مطالعه‌ای شبیه‌سازی بر روی داده‌های عاری از نقاط دورافُتاده[۲۷] و داده‌های آلوده شده[۲۸] انجام می‌شود. یکی از معیارهای مناسب برای تحقق این امر، استفاده از مقدار زاویه‌ای است که کرزانوسکی (۱۹۷۹) به‌صورت زیر تعریف کرده است:

فرض کنید زیرفضای پدید آمده توسط $k$ ویژه بردار نخست

[۲۲] Hyperparameter

[۲۳] Equidistant grid

[۲۴] Information criterion

[۲۵] The angle between subspaces

[۲۶] The average of total zero measure

[۲۷] Outliers-free data

[۲۸] Contaminated data

ماتریس $\Sigma$ به‌وسیلهٔ $E_k = Span\{e_1, ..., e_k\}$ تعریف شود که بردار $e_j$ ستون $j$اُم ماتریس $I_{p,k}$ است. در ماتریس نامبرده، $p$ نشان‌دهندهٔ تعداد متغیرها و $k$ نشان‌دهندهٔ تعداد مؤلفه‌های اصلی است. اکنون، بیشترین زاویه بین $E_k$ و زیرفضای پدید آمده توسط ستون‌های ماتریس بار برآورد شده ($P_{p,k}$ یا به‌طور خلاصه $P$) برای هر رویکرد از طریق رابطهٔ

$$\text{maxsub} = \arccos(\sqrt{\lambda_k})$$

محاسبه می‌شود که در آن، $\lambda_k$ کوچک‌ترین ویژه مقدار ماتریس $I'_{k,p}P_{p,k}P'_{k,p}I_{p,k}$ است. گفتنی است که این معیار، مقادیری بین ۰ تا $\frac{\pi}{2}$ را اختیار می‌کند اما به‌منظور یکسان‌سازی نتایج، مقادیر زوایا بر $\frac{\pi}{2}$ تقسیم خواهد شد تا مقادیر نهایی، اعدادی بین ۰ تا ۱ به دست آیند. بدیهی است که هر چه مقادیر حاصل به صفر نزدیک‌تر باشند، نشان‌دهندهٔ عملکرد مناسب رویکرد به‌کار رفته خواهد بود.

یکی از معیارهایی که به‌منظور مقایسه برآورد تُنُکی رویکردهای مطرح شده در این مقاله به‌کار گرفته می‌شود، معیار صفر بودن[۲۹] است ( هابرت و دیگران ۲۰۱۶). در واقع این معیار، ماتریس $P$ را به‌منظور تشخیص درستی ساختار تُنُکی آن مد نظر قرار می‌دهد. معیار حاضر با توجه به هر عضو ماتریس $P$، مقادیر صفر و یک را به‌صورت زیر اختیار می‌کند:

$$\text{مقدار برآورد شده و واقعی هر دو صفر یا هر دو غیر صفر باشند} \quad 1$$
$$\text{در غیر این صورت} \quad 0$$
$$= \text{معیار صفر بودن}$$

آن‌گاه، به میانگین تعداد صفرها روی تمامی عناصر ماتریس بار برآورد شده $P$ و تعداد تکرارهای شبیه‌سازی، معیار کل صفرها[۳۰] می‌گویند.

در شبیه‌سازی پیش‌رو، ساختار ماتریس همبستگی برای تولید ماتریسی با ویژه بردارهای تُنُک نقش بسزایی دارد. از این رو، ماتریس همبستگی با $k+1$ گروه از متغیرها طراحی می‌شود به‌گونه‌ای که بین متغیرهای گروه‌های مختلف، همبستگی وجود نداشته باشد. هر کدام از $k$ گروه نخست این ماتریس، شامل حد اقل ۴ متغیر ($b \geq 4$) هستند. همچنین، همبستگی میان متغیرهای گروه‌های $1, 2, ..., k$ به‌ترتیب برابر $a_1, a_2, ..., a_k$ در نظر گرفته می‌شوند. شایان ذکر است که گروه ($k+1$)اُم این ماتریس شامل $p-(k \times b)$ متغیر باقی‌مانده است و متغیرهای آن ناهمبسته فرض می‌شوند ($a_{k+1} = 0$). حال، ویژه بردار تُنُک $i$اُم این ماتریس برای $i = 1, 2, ..., k$ به‌گونه‌ای حاصل می‌شود که از عضو $(b \times (i-1)+1)$اُم تا عضو $(b \times i)$اُم آن مقدار $(-\frac{1}{\sqrt{b}})$ و بقیه عناصر، مقدار صفر را اختیار می‌کنند. به‌طور دقیق‌تر، هنگامی که نامساوی $a_1 > a_2 > ... > a_k$ برقرار باشد، $k$ ویژه بردار تُنُک ماتریس همبستگی مذکور به‌ترتیب برابر $p_1 = -\frac{1}{\sqrt{b}}q_1, p_2 = -\frac{1}{\sqrt{b}}q_2, ..., p_k = -\frac{1}{\sqrt{b}}q_k$ خواهند بود که در آن، $q_i \in R^p (i = 1, 2, ..., k)$ برداری است که $b$ عنصر $i$اُم آن مقدار یک و بقیه‌اش صفر هستند. پس از تولید ماتریس همبستگی ($R$)، این ماتریس با استفاده از تساوی $\Sigma = V^{\frac{1}{2}}RV^{\frac{1}{2}}$ به ماتریس کواریانس تبدیل می‌شود که $V$ نشان‌دهندهٔ ماتریس قطری شامل واریانس متغیرها است.

اکنون $n$ مشاهده از توزیع نرمال $p$-متغیره با میانگین $\mathbf{0}_p$ و ماتریس کواریانس $\Sigma$ تولید می‌شود. همچنین برای ایجاد ساختار تُنُک در داده‌ها، عبارت‌های نوفه[۳۱] که به‌صورت نرمال توزیع شده‌اند به هر یک از $p$ متغیر اضافه می‌شود. بنا بر این، مجموعه‌داده‌ها به‌صورت $X = X_u + X_{noise}$ تولید می‌شوند که $X_u \sim N_p(\mathbf{0}, \Sigma)$ و $X_{noise} \sim N_p(\mathbf{0}, I_p)$. سرانجام، برای داشتن توجیه مناسب در استفاده از رویکردهای اُستوار، به‌طور تصادفی $\varepsilon$٪(۱۰۰) از نقاط داده با نقاط دورافتاده جایگزین می‌شوند.

برای انجام شبیه‌سازی پیش‌رو، مجموعه‌ای با ده متغیر ($p = 10$) در نظر گرفته می‌شود. به علاوه، در تمامی مراحل شبیه‌سازی فرض می‌شود $k = 2$، $b = 4$، $a_1 = 0.9$ و $a_2 = 0.5$. بنا به مقادیر مفروض همبستگی، ویژه مقادیر ماتریس $R$ به‌ترتیب نزولی برابر با $3.7, 2.5, 1, 1, 0.5, 0.5, 0.5, 0.1, 0.1, 0.1$ خواهند بود. بنا بر این، دو ویژه بردار نخست ماتریس $R$ به‌صورت $p_1 = (-\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, 0, 0, 0, 0, 0, 0)'$ و $p_2$

[۲۹] Zero measure

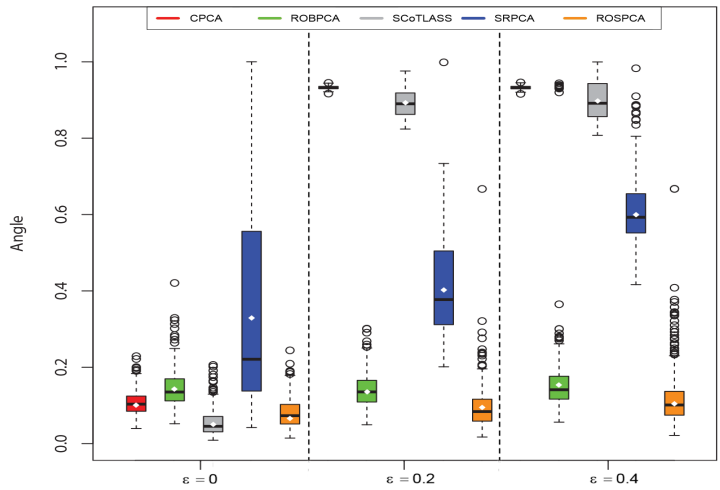
[۳۰] Total zero measure

[۳۱] Noise terms

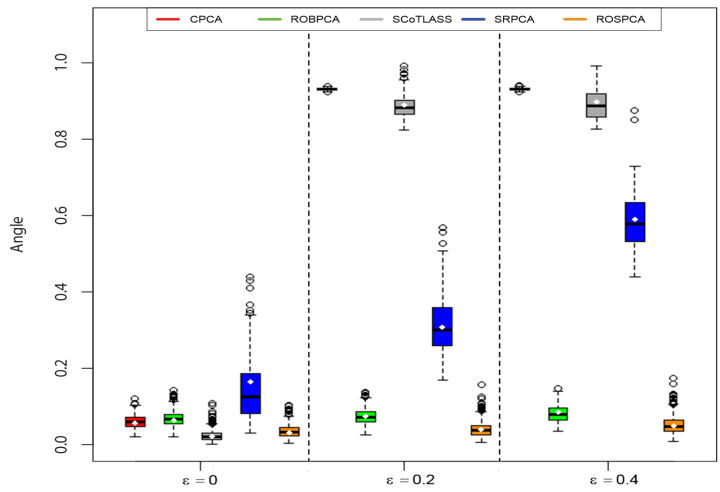
$(۰,۰,۰,۰,-\frac{۱}{۲},-\frac{۱}{۲},-\frac{۱}{۲},-\frac{۱}{۲},۰,۰)'$ هستند. اکنون برای تولید ماتریس کواریانس، ماتریس قطری $V$ به‌صورت $V = \text{diag}\{۱۰۰,۱۰۰,۱۰۰,۱۰۰,۲۵,۲۵,۲۵,۲۵,۴,۴\}$ در نظر گرفته می‌شود که نشان‌دهندهٔ برابری واریانس‌های متغیرهای هر گروه است. در نهایت، برای حصول مقدار بهینه پارامتر تُنُکی در رویکردهای، SCoTLASS SRPCA و، ROSPCA مقادیر $\lambda$ یکی از مقادیر $\{۰,۰/۰۲,...,۲/۴۸,۲/۵\}$ فرض می‌شوند. شایان ذکر است که پارامتر $\alpha$ برای رویکردهای ROBPCA و ROSPCA برابر ۰/۵ در نظر گرفته می‌شود.

سرانجام، به‌منظور ارزیابی و مقایسه عملکرد رویکرد ROSPCA با سایر رویکردها، ۵۰۰ مجموعه‌داده طبق طرح شبیه‌سازی تشریح شده با نسبت‌های مختلف دورافتادگی $(\varepsilon = ۰,۰/۲,۰/۴)$ برای هر رویکرد تولید و معیارهای مطرح شده محاسبه می‌شوند. نتایج شبیه‌سازی به شرحی که در ادامه می‌آیند، هست.

در ابتدا، نتایج حاصل از رویکردهای مطرح شده به‌صورت نمودارهای جعبه‌ای گزارش می‌شوند که این نمودارها بر مبنای بیشترین مقدار زاویه به‌دست‌آمده از هر رویکرد هستند.




(الف)




(ب)

**شکل ۱.** مقادیر زاویه به‌دست‌آمده از رویکردهای، CPCA، ROBPCA، SCoTLASS SRPCA و ROSPCA به‌ازای $p = ۱۰$ و $\varepsilon = ۰,۰/۲,۰/۴$ برای (الف) $n = ۱۰۰$ و (ب) $n = ۳۰۰$.

**جدول ۱.** میانگین کل صفرهای حاصل از رویکردهای، CPCA، ROBPCA، SCoTLASS SRPCA و ROSPCA به‌ازای $p = ۱۰$ و مقادیر متفاوت $n$ و $\varepsilon$.

| | $n = ۳۰۰$ | | | $n = ۱۰۰$ | | |
|---|---|---|---|---|---|---|
| | $\varepsilon = ۰/۴$ | $\varepsilon = ۰/۲$ | $\varepsilon = ۰$ | $\varepsilon = ۰/۴$ | $\varepsilon = ۰/۲$ | $\varepsilon = ۰$ |
| CPCA | ۰/۴ | ۰/۴ | ۰/۴ | ۰/۴ | ۰/۴ | ۰/۴ |
| ROBPCA | ۰/۴ | ۰/۴ | ۰/۴ | ۰/۴ | ۰/۴ | ۰/۴ |
| SCoTLASS | ۰/۳۸ | ۰/۳۹ | ۰/۹۶ | ۰/۳۹ | ۰/۳۶ | ۰/۹۶ |
| SRPCA | ۰/۴۲ | ۰/۴۶ | ۰/۶۹ | ۰/۴۳ | ۰/۵ | ۰/۷۱ |
| ROSPCA | ۰/۸۷ | ۰/۹۱ | ۰/۹۵ | ۰/۸۷ | ۰/۸۹ | ۰/۹۱ |

با نگاهی به شکل ۱ مشاهده می‌شود که با افزایش تعداد مشاهدات نمونه ($n$)، اُریبی برآوردها و پراکندگی مقادیر زاویه کاهش یافته‌اند. نمودارهای جعبه‌ای رویکرد CPCA نشان می‌دهند که این روش در داده‌های عاری از نقاط دورافتاده ($\varepsilon = ۰$) همانند رویکردهای SCoTLASS و ROSPCA مدل مناسبی است. اما در صورت اعمال آلودگی ($\varepsilon = ۰/۲, ۰/۴$) و حضور نقاط دورافتاده، مقادیر زاویه متناظر با آن افزایش پیدا می‌کنند. نمودارهای رویکرد ROBPCA نیز در دو حالت $n$ نشان می‌دهند که با افزایش نرخ آلودگی، مقادیر زاویه مقداری افزایش می‌یابند. به‌عبارتی، این رویکرد نسبت به رویکردهای، CPCA SCoTLASS و SRPCA در حضور نقاط دورافتاده عملکرد بهتری دارد اما برای شناسایی ساختار تُنُک داده‌ها مناسب نیست. گفتنی است که برای حالت $\varepsilon = ۰$، بهترین نتیجه ممکن مربوط به رویکرد SCoTLASS است. اما این رویکرد نیز با حضور نقاط دورافتاده عملکرد ضعیفی از خود نشان می‌دهد. رویکرد ترکیبی SRPCA نیز بر خلاف رویکرد، ROSPCA در حالت‌های مختلف $n$ و نرخ‌های آلودگی، مقادیر زاویه نسبتاً بزرگی را تولید می‌کند و از اُریبی بالایی نیز برخوردار است. اما ملاحظه می‌شود که رویکرد جدید ROSPCA در تمامی حالات $n$ و سطوح آلودگی حتی زمانی که $\varepsilon = ۰/۴$، اُریبی کمتری را از خود نشان می‌دهد. به علاوه، می‌توان گفت که در حضور و یا عدم حضور نقاط دورافتاده، رویکرد ROSPCA عملکرد بهتری نسبت به سایر رویکردهای منتسب به PCA دارد.

اکنون صحت تُنُکی رویکردهای مذکور با استفاده از معیار صفر بودن بررسی می‌شود. با نگاهی به جدول ۱ می‌توان گفت که رویکرد SCoTLASS در حالات مختلف $n$ زمانی که مجموعه‌داده‌ها عاری از نقاط دورافتاده ($\varepsilon = ۰$) باشند، به خوبی ماتریس بار تُنُک $P$ را برآورد می‌کند در حالی که با افزایش نرخ آلودگی، تُنُکی ماتریس بار به خوبی تضمین نمی‌شود. همچنین، میانگین کل صفرهای مربوط به رویکرد SRPCA نشان می‌دهد که این روش عملکرد ضعیف‌تری نسبت به سایر رویکردها دارد. با مشاهده نتایج رویکرد ROSPCA مشخص است که این روش ترکیبی به‌درستی ساختار تُنُک ماتریس $P$ را در تمامی حالات حفظ می‌کند. به بیانی دیگر، رویکرد ROSPCA حتی با وجود تعداد زیادی نقاط دورافتاده در مجموعه‌داده‌ها، ماتریس بار تُنُک $P$ را به خوبی برآورد می‌کند. قابل ذکر است که رویکردهای CPCA و ROBPCA با توجه به ماهیت اصلی خود به سختی بارهای صفر تولید می‌کنند. لذا، میانگین کل صفرهای آنها به‌طور ثابت برابر مقدار ۰/۴ که درصدی از مقادیر ورودی غیر صفر در ماتریس $P$ است، خواهد بود.

نتیجه کلّی حاصل از انجام شبیه‌سازی نشان می‌دهد که رویکرد ROSPCA نه تنها برآوردهای اُستواری ارائه می‌کند بلکه به‌منظور شناسایی ساختار تُنُک داده‌ها از سایر رویکردهای PCA کارایی به مراتب بهتری دارد.

## ۲.۴ تحلیل مثال واقعی

برای پیاده‌سازی رویکرد، ROSPCA داده‌های جرم و جنایت مربوط به ایالت‌های آمریکا که در سال ۱۹۸۶ گزارش شده است، تحلیل می‌شود. این مجموعه‌داده در پی تحقیق به‌منظور بررسی نرخ انواع متفاوت جرم و جنایت بر حسب ۱۰۰۰۰۰ نفر ساکنین ۵۰ ایالت آمریکا گردآوری شده است (اوریت و هاترن، ۲۰۱۱). در این مجموعه‌داده، ۷ نوع متفاوت از جرائم شامل دو

گروه جرائم خشونت‌آمیز (صدمات به اشخاص) و جرائم اموالی (صدمات به منازل مسکونی) ثبت شده است. قابل ذکر است که جرائم خشونت‌آمیز شامل قتل و کشتار $(X_۱ = Murder)$، تجاوز به عنف $(X_۲ = Rape)$، سرقت $(X_۳ = Robbery)$ و یورش $(X_۴ = Assualt)$ و جرائم اموالی شامل ورود غیر قانونی به خانه و دزدی از آن $(X_۵ = Burglary)$، دستبرد به اموال شخصی دیگران $(X_۶ = Theft)$ و سرقت وسایل نقلیه موتوری $(X_۷ = Vehicle)$ هستند.

از آن‌جا که برآوردگر MCD در مقابل نقاط دورافتاده اُستوار است، فاصله ماهالانوبیس مبتنی بر آن می‌تواند یکی از معیارهای مناسب برای تشخیص دقیق چنین نقاطی باشد (هابرت و دی بروین، ۲۰۱۰). شکل ۲، نمودار ریشه توان دوم فاصله اُستوار بر حسب ایالت‌های آمریکا است که در آن، نقاط موجود در بالای خط برش (خط آستانه) نشان‌دهندهٔ نقاط دورافتاده هستند. لذا در این حالت، استفاده از رویکرد PCA به‌منظور کشف منابع اصلی تغییرات نتایجی غیر معتبر به همراه خواهد داشت. از این رو، استفاده از رویکردهای اُستوار مطرح شده برای اخذ نتایج مورد اعتماد، مفید فایده خواهد بود. اما از آن‌جا که حصول مؤلفه‌های اصلی تعبیرپذیر نیز یکی از اهداف اصلی محقق است، پیاده‌سازی رویکرد ترکیبی اُستوار و تُنُک ROSPCA بر روی داده‌های حاضر دنبال می‌شود.

با توجه به جدول ۲ که نشان‌دهندهٔ درصد واریانس تجمعی مؤلفه‌های اصلی است، می‌توان تعداد مؤلفه‌های مؤثر در اجرای رویکرد را انتخاب کرد. اگر مبنای بیان واریانس تجمعی با حد اقل ۹۰٪ به‌عنوان حد اقل مورد نیاز باشد، واضح است که تعداد سه مؤلفه اصلی $(k = ۳)$ برای به‌کارگیری رویکرد ROSPCA مناسب است.

برای تعیین مقدار بهینه $\lambda$، مقادیر $\lambda$ با فاصله ۰٫۰۲ به‌صورت $\{۰, ۰٫۰۲, ۰٫۰۴, ..., ۱٫۹۶, ۱٫۹۸, ۲\}$ در نظر گرفته شدند. شکل ۳، نمودار مقادیر رابطهٔ (۲) بر حسب ۵۰ مقدار $\lambda$ مفروض است که نشان می‌دهد به‌ازای $\lambda = ۰٫۵۴$، معیار BIC کمترین مقدار خود (۲٫۸۰۱−) را اختیار می‌کند. لذا، مقدار پارامتر تُنُکی برابر ۰٫۵۴ انتخاب می‌شود. اکنون با توجه به مقادیر تعیین شده، رویکرد ROSPCA اجرا و نتایج آن بیان می‌شود.

شکل ۴، نمودار تشخیصی مربوط به داده‌های جرم و جنایت در ایالت‌های آمریکا را پس از اعمال رویکرد ROSPCA نشان می‌دهد. با توجه به این شکل مشاهده می‌شود که گروهی از مشاهدات، مقادیر OD بالاتری از خط برش دارند که تحت عنوان نقاط دورافتاده یاد می‌شوند. لذا، می‌توان دریافت که رویکرد نامبرده می‌تواند نقاط دورافتاده موجود در مجموعه‌داده‌ها را به‌درستی کشف و شناسایی کند.

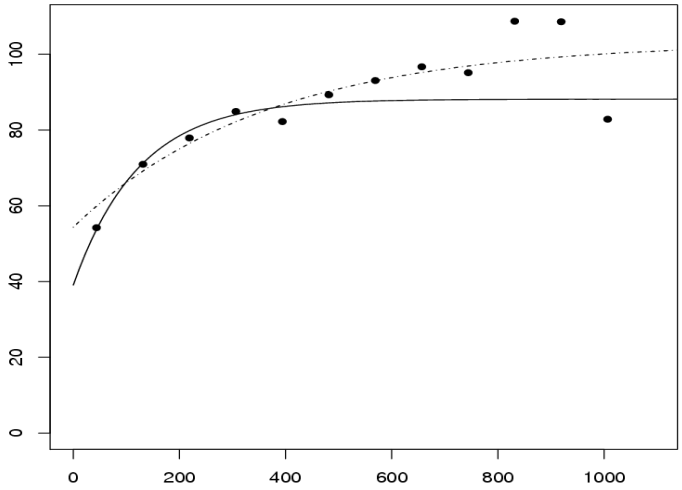
علاوه بر این، نمودار مقادیر بارهای حاصل از کاربست رویکرد ROSPCA بر روی مجموعه‌داده تحت بررسی در شکل ۵ نشان داده شده است. با توجه به این شکل، مشاهده می‌شود که تمامی بارهای مؤلفه اصلی نخست، غیر صفر هستند. به‌عبارت دیگر، به نظر می‌رسد که تمامی متغیرهای موجود در آن تقریباً به یک میزان نقش ایفا می‌کنند و نمی‌توان نقش متغیری را نسبت به سایر متغیرها بیشتر یا کمتر دانست. همچنین، با نگاهی به مقادیر بارهای مؤلفه اصلی دوم نیز مشخص است که بارهای مربوط به متغیرهای $X_۲$ و $X_۴$ صفر بوده و پنج متغیر دیگر دارای ضرایبی غیر صفر هستند. در مؤلفه اصلی سوم نیز، متغیرهای $X_۲$، $X_۵$ و $X_۶$ دارای بارهای صفر بوده و چهار متغیر باقی‌مانده شامل بارهای غیر صفر هستند. بنا بر این ملاحظه می‌شود که تُنُکی رویکرد ROSPCA در این مسئله به خوبی لحاظ شده است.

نتایج حاصل از اعمال رویکرد ROSPCA بر روی مجموعه‌داده جرم و جنایت نشان داد که این ابزار توانست علاوه بر کاهش بُعد و توصیف منابع اصلی تغییرات موجود در ماهیت داده‌ها که از اهداف اصلی رویکردهای متفاوت PCA است، به‌درستی نقاط دورافتاده موجود در داده‌ها را کشف و ساختار تُنُک مناسبی ارائه کند.

## بحث و نتیجه‌گیری

در مقاله حاضر، چند دیدگاه متفاوت غلبه بر مشکلات رویکرد PCA معمولی مد نظر قرار گرفت. به‌طور دقیق‌تر، رویکردهای تحلیل مؤلفه‌های اصلی اُستوار برای حل مشکلات ناشی از

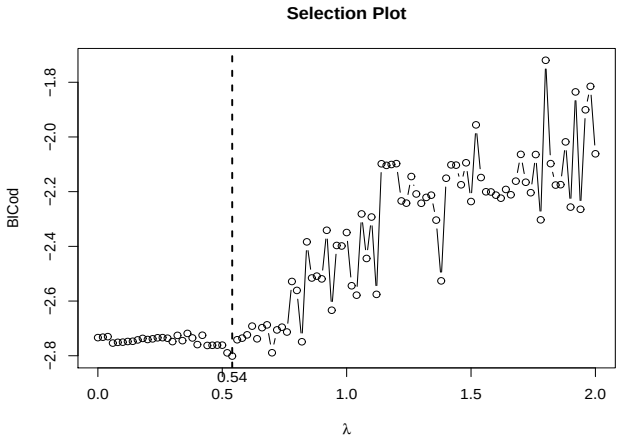
حضور نقاط دورافتاده و رویکردهای تُنُک‌سازی به‌منظور افزایش تعبیرپذیری از جمله روش‌های مفید هستند که مورد بحث قرار گرفت. وجود این دو مشکل به‌صورت هم‌زمان در بسیاری از مسائل آمار کاربردی باعث شد، رویکرد جدیدی برای تحلیل مؤلفه‌های اصلی اُستوار و تُنُک تحت عنوان رویکرد ROSPCA پا به عرصه وجود بگذارد که هدف آن به دست آوردن مؤلفه‌های اصلی اُستوار با مقادیر بار تُنُک بود. نتایج بررسی و ارزیابی شبیه‌سازی نیز نشان داد که رویکرد ROSPCA به مراتب عملکرد بهتری نسبت به سایر رویکردهای متفاوت PCA دارد.



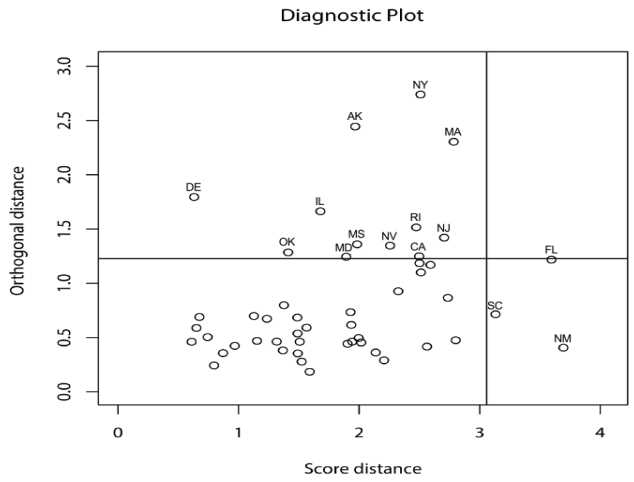

**شکل ۲.** نمودار ریشه توان دوم فاصله اُستوار بر حسب ایالت‌های آمریکا

**جدول ۲.** تغییرات تجمعی مؤلفه‌های اصلی (به درصد)

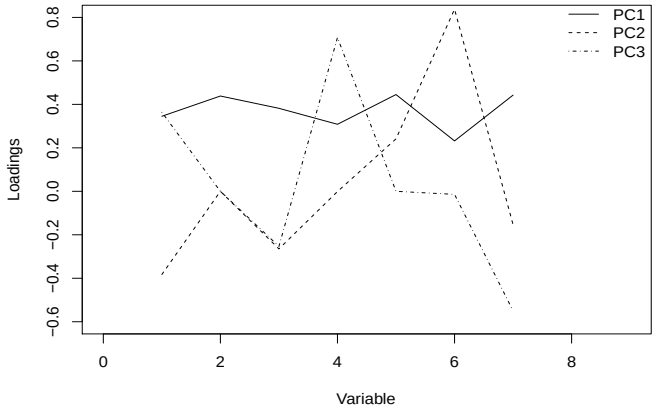
| PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|
| ۶۹٫۶ | ۸۸٫۵ | ۹۵٫۰ | ۹۷٫۰۷ | ۹۸٫۵۳ | ۹۹٫۶۶ | ۱۰۰ |




**شکل ۳.** نمودار مقادیر معیار IC بر حسب مقادیر مفروض $\lambda$.نمودار ریشه توان دوم فاصله اُستوار بر حسب ایالت‌های آمریکا




**شکل ۴.** نمودار تشخیصی مربوط به رویکرد ROSPCA ($k = ۳$، $\alpha = ۰/۵$ و $\lambda = ۰/۵۴$).




**شکل ۵.** نمودار مقادیر بار حاصل از رویکرد ROSPCA ($k = ۳$، $\lambda = ۰/۵۴$ و $\alpha = ۰/۵$).

# مراجع

[1] Croux, C., Filzmoser, P.. and Fritz, H. (2013). Robust Sparse Principal Component Analysis. *Technometrics*, **55**, 202-214.

[2] Croux, C., and Haesbroeck, G. (1999). Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis*, **71**, 161-190.

[3] Everitt, B., and Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*, Springer, New York.

[4] Huber, P. J. (1985). Projection Pursuit. *The Annals of Statistics*, **13**, 435-475.

[5] Hubert, M., and Debruyne, M. (2010). Minimum Covariance Determinant. *Computational Statistics*, **2**, 36-43.

[6] Hubert, M., Reynkens, T., Schmitt, E., and Verdonck, T. (2016). Sparse PCA for High-Dimensional Data with Outliers. *Technometrics*, **58**, 424-434.

[7] Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, **47**, 64-79.

[8] Jolliffe, I. (2002). *Principal Component Analysis*, 2nd Edition, Springer, New York.

[9] Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**, 531-547.

[10] Krzanowski, W. J. (1979). Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association*, **74**, 703-707.

[11] Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications*, **8**, 283-297.

[12] Rousseeuw, P. J., and Driessen, K. V. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**, 212-223.

[13] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Methodological*, **58**, 267-288.