# شناسایی نقاط دورافتاده چندمتغیره با استفاده از فواصل استوار

اباذر خلجی، زهرا شاه‌محمدی
گروه آمار، دانشگاه اصفهان

**چکیده**

برای شناسایی نقاط دورافتاده یک‌متغیره روش‌هایی وجود دارد که در حالت چندمتغیره کارامد نیستند. علاوه بر این زمانی که چندین مشاهده دورافتاده در مجموعه داده‌های چندمتغیره وجود دارد، شناسایی نقاط دورافتاده چندمتغیره امری دشوار و نیازمند روش‌های محاسباتی می‌باشد. روش‌های شناسایی کلاسیک اغلب قادر به شناسایی این نقاط نیستند. زیرا بر پایه بردار میانگین و ماتریس کوواریانس آلوده به شناسایی نقاط دورافتاده می‌پردازند. به همین دلیل با مساله درون‌آوری مواجه می‌شوند. برای مقابله با درون‌آوری فواصل باید بر اساس برآوردگرهای استوار محاسبه شوند. در این مقاله از برآوردگرهای کمینه دترمینان ماتریس کوواریانس و بیضی‌گون کمینه حجم که دارای بیشینه فروریختگی می‌باشند در محاسبه فواصل بهره جسته‌ایم.

**واژه‌های کلیدی:** نقاط دورافتاده، فاصله ماهالانوبیس، فاصله استوار.

## ۱ مقدمه

نقاط دورافتاده[1] مقادیری در مجموعه داده‌ها هستند که غالباً از الگوی اکثر مشاهدات پیروی نمی‌کنند. هاوکینز [۴] نقاط دورافتاده را این‌گونه تعریف می‌کند:

«نقاط دور افتاده مشاهداتی هستند که به قدری دورتر از سایر مشاهدات قرار می‌گیرند که گویی با مکانیسم دیگری تولید شده‌اند.»

گاهی نقاط در اثر خطاهای اندازه‌گیری، ثبت نادرست یا برخی عوامل دیگر آلوده می‌شوند. بعضی از محققین عقیده دارند که در یک پژوهش معمولاً بیش از ۱۰ درصد مشاهدات دارای خطاهای اندازه‌گیری و جمع‌آوری هستند (رنچر، [۷]). اگر آلودگی مشاهدات ناشی از خطاهای اندازه‌گیری و ثبتی بوده و امکان دسترسی مجدد برای اصلاح آن وجود نداشته باشد، باید آن‌ها را حذف کرد، اما در غیر این صورت چون حاوی اطلاعات در مورد نمونه هستند حذف آن‌ها جایز نیست.

وجود نقاط دورافتاده در نمونه‌های یک یا چندمتغیره باعث تخریب برآوردگر (بردار) مکانی و (ماتریس) پراکندگی می‌شوند، توزیع داده‌ها را تغییر داده و منجر به توزیع‌هایی با دم‌های سنگین می‌شوند، میزان تغییرات را افزایش داده و همبستگی بین متغیرها را در هم می‌ریزند. بنابراین در مدل سازی آماری و تحلیل داده‌ها بررسی

[1] Outliers

وجود نقاط دورافتاده در مجموعه داده‌ها از اهمیت بالایی برخوردار است. در حالت یک‌بعدی (تک متغیره) امکان شناسایی نقاط دورافتاده از روی نمودار پراکنش داده‌ها وجود دارد و در حالت دو و سه متغیره نیز اگر نمودارهای پراکنش را در فضاهای دو یا سه بعدی ترسیم نماییم، احتمال شناسایی نقاط دورافتاده وجود دارد ولی لزوماً همیشه این کار شدنی نیست. بنابراین در داده‌های چندمتغیره نیازمند روش‌های دیگری می‌باشیم.

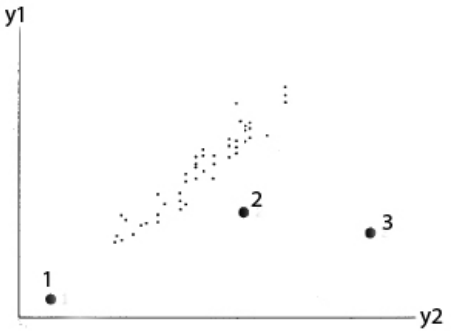
در داده‌های چندمتغیره مسئله شناسایی نقاط دورافتاده از اهمیت و دشواری‌های بیشتری در مقایسه با حالت یک‌متغیره برخوردار است، زیرا

۱. برای $p > ۲$ نمودار پراکنش داده‌ها قادر به شناسایی نقاط دورافتاده نمی‌باشد.

۲. هر مشاهده به‌صورت یک بردار بوده و ممکن است یک خطای (اندازه‌گیریی، ثبتی و ...) بزرگ در یکی از اعضای بردار یا خطاهای کوچک در چندین عضو بردار رخ دهد.

۳. داده‌های چندمتغیره را نمی‌توان مانند داده‌های تک متغیره به‌صورت صعودی مرتب کرد و مقادیر کوچک و بزرگی را که در دو انتها قرار می‌گیرند مشخص نمود.

۴. نقاط دورافتاده چندمتغیره ممکن است به واسطه تغییر در میانگین، واریانس یا همبستگی ایجاد شده باشند. برای مثال در شکل ۱ که از کتاب رنچر [۷] اقتباس شده، مشاهده ۱ با تغییر اندک میانگین و واریانس هر دو متغیر $y_۱$ و $y_۲$ به وجود آمده است، مشاهده ۲ به واسطه تغییر در همبستگی بین متغیرها ایجاد شده و مشاهده ۳ با تغییر میانگین، واریانس و همبستگی بین متغیرها به وجود آمده است.




شکل ۱: نمایش سه نوع آلودگی در نمونه دومتغیره

**تعریف۱** یک بیضی‌گون در فضای $p$ بعدی به‌صورت

$$(\underline{x}-\underline{b})^T A^{-۱}(\underline{x}-\underline{b}) = d^۲,$$

تعریف می‌شود، که در آن $p$ نشان دهنده محور مختصات، $A$ یک ماتریس $p\times p$ معین مثبت (و درکاربردهای آماری متقارن)، $\underline{b}$ مرکز بیضی‌گون و $d^۲$ یک عدد حقیقی مثبت می‌باشد.

**تعریف۲** اگر $\overline{\underline{x}}$ و $S$ به ترتیب بردار میانگین و ماتریس کوواریانس یک نمونه $n$ تایی باشد آنگاه فاصله

$$MD = \sqrt{(\underline{x}-\overline{\underline{x}})^T S^{-۱}(\underline{x}-\overline{\underline{x}})},$$

یک فاصله ماهالانوبیس[۲] برای $\underline{x}$ است.

در واقع یک بیضی‌گون مشخص کننده مجموعه تمامی نقاطی از فضای $p$ بعدی است که مربع فاصله ماهالانوبیسی آن نقاط تا مرکز $\overline{\underline{x}}$ برابر مقدار ثابت $d$ می‌باشد.

[۲] Mahalanobis Distance

## ۲ شناسایی به‌وسیله مربع فاصله ماهالانوبیس

می‌دانیم مشاهداتی که از انبوهه[۳] اصلی مشاهدات دور هستند دورافتاده تلقی می‌شوند. اگر توزیع مولد داده‌های نمونه نرمال چند متغیره باشد، در یک نمودار پراکنش چند بعدی انبوهه داده‌ها به شکل بیضی‌گون است. بیضی‌گونی که در مرکز آن فشردگی داده‌ها زیاد و هر چه از مرکز دور شویم فشردگی داده‌ها کمتر می‌شود. بنابراین اگر تعداد متغیرها ۱، ۲ و یا ۳ باشد با رسم نمودار پراکنش دوری بعضی مشاهدات از انبوهه اصلی به‌صورت بصری قابل مشاهده است. در حالت سه بعدی با چرخاندن نمودار، انبوهه بیضی‌گون و نقاط دور از آن انبوهه آشکار می‌شوند. در حالت $p > ۳$ این دوری قابل رویت نیست و لذا باید به طریق قابل اعتمادی اندازه‌گیری شود. روش‌های شهودی در شناسایی نقاط دورافتاده زمانی که تعداد متغیرها بیشتر از ۳ باشد غیر قابل اعتماد یا امکان ناپذیر هستند. از این رو به کمک روش‌های دیگر که غالباً محاسباتی هستند به شناسایی نقاط دورافتاده می‌پردازند. اولین روش که به ذهن خطور می‌کند، استفاده از مربع فاصله ماهالانوبیس[۴] می‌باشد که برای مشاهده $i$ ام به‌صورت

$$MSD_i = (\underline{x}_i - \overline{\underline{x}})^T S^{-۱} (\underline{x}_i - \overline{\underline{x}}), \qquad (۱)$$

تعریف می‌شود که $\overline{\underline{x}}$ میانگین حسابی و $S$ ماتریس کوواریانس نمونه‌ای مجموعه دادهای $X$ می‌باشد. طبیعی است که مشاهداتی که دارای فواصل بزرگ می‌باشند به عنوان نقاط دور افتاده شناسایی شوند. اما برای رسیدن به نتایج قابل اعتماد نیازمند مقدار برشی[۵] هستیم تا مقادیر $MSD_i$ را با آن بسنجیم.

اگر بردار میانگین و ماتریس کوواریانس جامعه یعنی $\underline{\mu}$ و $\Sigma$ معلوم باشند آن‌گاه $MSD_i$ دارای توزیع $\chi^۲_p$ (مربع کای با $p$ درجه آزادی) می‌باشد و اگر از برآوردگرهای کلاسیک $\overline{\underline{x}}$ و $S$ در محاسبه فاصله ماهالانوبیس استفاده کنیم، فاصله دارای توزیع مجانبی $\chi^۲_p$ خواهد بود، [۶]. بنابراین اگر مربع فاصله ماهالانوبیس مشاهده $i$ام از مقدار برشی $\chi^۲_{\%۲۵,p}$ بزرگتر باشد آن‌گاه مشاهده $i$ام به عنوان مشاهده دورافتاده شناسایی می‌شود.

درواقع $MSD_i$ فاصله مشاهده $\underline{x}_i$ را تا مرکز همه داده‌ها نه مرکز انبوهه اصلی اندازه‌گیری می‌کند. به دلیل اینکه نقاط دور افتاده در محاسبه $\overline{\underline{x}}$ و $S$ نیز حضور دارند دیگر با فاصله‌های بزرگ مواجه نخواهیم بود و لذا استفاده از معیار $MSD$ درون‌آوری[۶] را در پی خواهد داشت. به‌عبارت دیگر زمانی که با چندین داده دورافتاده مواجه هستیم این مشاهدات آلوده بر میانگین تاثیر می‌گذارند و باعث می‌شوند که میانگین به مشاهدات دورافتاده نزدیک شود. نزدیکی میانگین به مشاهدات دورافتاده و اثر این مشاهدات بر ماتریس کوواریانس باعث کاهش فاصله ماهالانوبیس شده و در نهایت از شناسایی نقاط دورافتاده جلوگیری می‌کند. با این ویژگی‌ها چون برآوردگرهای کلاسیک تحت تاثیر نقاط دورافتاده هستند قادر به شناسایی نقاط پرت نخواهیم بود، به این ترتیب مقصود از درون‌آوری پنهان شدن مشاهدات دورافتاده به واسطه برآوردگرهای مکانی و پراگندگی آلوده می‌باشد. برای کنکاش بیشتر در مورد درون‌آوری به خلجی و خردمندنیا [۱] رجوع کنید.

[۳] Cloud
[۴] Mahalanobis Squared Distance
[۵] Cutoff Value
[۶] Masking

## ۳ شناسایی به‌وسیله مربع فاصله استوار

$\overline{\underline{x}}$ مرکز همه داده‌ها است ولی مرکز انبوهه اصلی داده‌ها نیست. به طریق مشابه $S$ نیز اطلاعات پراکندگی همه داده‌ها را در بر دارد و فاقد اطلاعات انبوهه اصلی می‌باشد. بنابراین با توجه به مباحث بخش قبل معقول به نظر می‌رسد که به جای برآوردگرهای کلاسیک $\overline{\underline{x}}$ و $S$ از برآوردگرهای استوار[۷] استفاده شود که در برابر نقاط دورافتاده مقاوم بوده و حاوی اطلاعات انبوهه اصلی داده‌ها می‌باشند. در این مقاله براوردگرهای استوار بردار میانگین و ماتریس کوواریانس را به ترتیب با $\widetilde{\underline{\mu}}$ و $\widetilde{\Sigma}$ نشان می‌دهیم.

کمپل [۲] پیشنهاد کرد که برآوردگرهای-$M$ را جایگزین $\overline{\underline{x}}$ و $S$ کنیم که بهبود حائز اهمیتی حاصل شد. اما نقطه فروریختگی این برآوردگرها حداکثر برابر $\frac{۱}{p+۱}$ می‌باشد (روسو[۸]) که با افزایش بعد داده‌ها به شدت دچار فروریختگی می‌شود.

در پژوهش‌های بعد برآوردگرهای بردار مکانی و ماتریس پراکندگی که دارای بیشینه فروریختگی بودند جایگزین $\overline{\underline{x}}$ و $S$ شدند. برای اولین بار استاهل [۱۰] و دونوهو و هابر [۳] چنین برآوردگرهایی را توصیه کردند. در این بخش ما از برآوردگرهای کمینه دترمینان ماتریس کوواریانس ($MCD$)[۸] که دارای نقطه فروریختگی ۵۰ درصدی می‌باشند، استفاده می‌کنیم. ایده اولیه این روش متعلق به روسو و زومرن [۹] می‌باشد. برآورد استوار مبتنی بر روش $MCD$ را با $\widetilde{\underline{\mu}}_{(MCD)}$ و $\widetilde{\Sigma}_{(MCD)}$ نشان می‌دهیم. به این ترتیب مربع فاصله استوار[۹] مشاهده $i$ ام عبارت است از

$$RSD_i = (\underline{x}_i - \widetilde{\underline{\mu}}_{(MCD)})^T \widetilde{\Sigma}^{-۱}_{(MCD)} (\underline{x}_i - \widetilde{\underline{\mu}}_{(MCD)}).$$

روسو و زومرن [۹] پیشنهاد کردند از مقدار برشی $\chi^۲_{\alpha,p}$ استفاده شود و مقدار $\alpha = ۰/۰۲۵$ را به‌عنوان یک مقدار مرسوم معرفی کردند. بنابراین مشاهده $\underline{x}_i$ دورافتاده تلقی می‌شود اگر مربع فاصله استوار آن از مقدار برشی $\chi^۲_{۰/۰۲۵,p}$ بیشتر باشد.

با احتمال زیاد می‌توانیم امیدوار باشیم که برآوردهای $MCD$، پارامتر مکانی و پراکندگی را اندازه‌گیری می‌کند که حدود نیمی از داده‌های سالم را در بر دارد. بنابراین استفاده از برآوردگرهای استوار منجر به تولید فواصل استواری شده که تحت تاثیر نقاط دورافتاده نبوده و لذا استفاده از آنها درون‌آوری را از بین می‌برد. برای روشن شدن مطلب به مثال‌های زیر توجه کنید. لازم به ذکر است که در این مثال‌ها از فواصل (و نه از مربع آنها) و مقدار برشی $\sqrt{\chi^۲_{\alpha,p}}$ برای شناسایی نقاط دورافتاد استفاده شده است.

**مثال۱: داده‌های وزن مغز و بدن**

این مجموعه داده‌ها شامل وزن مغز و وزن بدن ۲۸ گونه جانوری است، که توسط دانشگاه ایلینوی جمع‌آوری شده است. وزن بدن جانوران بر حسب کیلوگرم و وزن مغز آنها بر حسب گرم محاسبه شده است. داده‌های خام تحت عنوان *Animals* در نرم‌افزار $R$ موجود می‌باشند. در این مثال با استفاده از برنامه‌ای که در پیوست ارائه

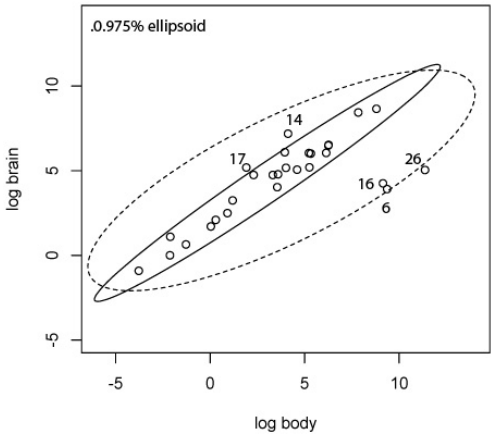
[۷] Robust
[۸] Minimum Covarianc Determinant
[۹] Robust Squared Distance

شده فواصل ماهالانوبیس $(MD)$ و فاصله استوار $(RD)$ را محاسبه کرده و نتایج را در جدول ۱ قرار داده‌ایم. با استفاده از مقدار برشی $\sqrt{\chi^2_{۰٫۵,۲}} = ۲٫۷۲$ و فاصله ماهالانوبیس مشاهدات تنها مشاهده ۲۶ که مربوط به نوعی دایناسور $(Brachiosaurus)$ می‌باشد به‌عنوان نقطه دورافتاده شناسایی می‌شود. اما با استفاده از فاصله استوار مشاهده‌های ۶, ۱۶ و ۲۶ که سه دایناسور با مغز کوچک و بدن سنگین هستند و مشاهدات ۱۴ و ۱۷ که مربوط به بشر و نوعی میمون است به‌عنوان داده دورافتاده شناسایی می‌شوند.

جدول ۱: فاصله ماهالانوبیس و فاصله استوار برای داده‌های وزن مغز و بدن

| $RD$ | $MD$ | مشاهده | $RD$ | $MD$ | مشاهده |
|---|---|---|---|---|---|
| ۱,۷۲ | ۱,۷۶ | ۱۵ | ۰,۸۶ | ۱,۰۱ | ۱ |
| ۹,۵۸ | ۲,۳۷ | ۱۶ | ۱,۴۸ | ۰,۷ | ۲ |
| ۳,۵۷ | ۱,۲۲ | ۱۷ | ۰,۲۴ | ۰,۳ | ۳ |
| ۱,۳۲ | ۰,۲ | ۱۸ | ۰,۵۲ | ۰,۳۸ | ۴ |
| ۱,۷۴ | ۱,۸۶ | ۱۹ | ۱,۱۱ | ۱,۱۵ | ۵ |
| ۲,۰۳ | ۲,۲۷ | ۲۰ | ۱۰,۶ | ۲,۶۴ | ۶ |
| ۰,۷۴ | ۰,۸۳ | ۲۱ | ۱,۸۱ | ۱,۷۱ | ۷ |
| ۰,۴۲ | ۰,۴۲ | ۲۲ | ۰,۶۹ | ۰,۷۱ | ۸ |
| ۰,۹۴ | ۰,۲۶ | ۲۳ | ۱,۰۵ | ۰,۸۶ | ۹ |
| ۲,۲۹ | ۱,۰۵ | ۲۴ | ۲,۰۸ | ۰,۸۰ | ۱۰ |
| ۱,۵۲ | ۱,۵۹ | ۲۵ | ۰,۸۹ | ۰,۶۹ | ۱۱ |
| ۱۱,۴۷ | ۲,۹۱ | ۲۶ | ۱,۰۴ | ۰,۸۷ | ۱۲ |
| ۲,۰۵ | ۱,۵۸ | ۲۷ | ۰,۶۸ | ۰,۶۸ | ۱۳ |
| ۱,۶۸ | ۰,۳۹ | ۲۸ | ۴,۲۶ | ۱,۷۲ | ۱۴ |

در شکل ۲ بیضی ۹۷/۵ درصد برای لگاریتم داده‌های وزن مغز و بدن ارائه شده است. بیضی‌گون ماهالانوبیس با خط چین و بیضی‌گون استوار با خط ممتد مشخص شده که داده‌های خارج از بیضی‌گون‌ها پرت تلقی می‌شوند. با توجه به مثال ۱ می‌توان نتیجه گرفت که فاصله ماهالانوبیس به علت درون‌آوری قادر به شناسایی نقاط دور افتاده نبوده که با استفاده از فاصله استوار این مشکل مرتفع می‌شود.




شکل ۲: بیضی‌گون ۹۷/۵ درصد برای لگاریتم داده‌های وزن مغز و بدن

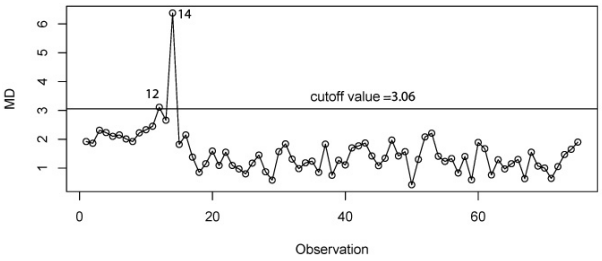
**نکته۱** شناسایی نقاط دورافتاده زمانی که $\frac{n}{p}$ کوچک باشد امری دشوار است. زیرا در این حالت نقاط کمتری در محاسبه $MCD$ شرکت دارند. با کمک شبیه سازی و به‌عنوان یک قاعده سرانگشتی توصیه می‌شود که اندازه نمونه حداقل باید ۵ برابر تعداد متغیرها باشد.

### مثال۲: داده‌های هاوکینز و همکاران

در این مثال با استفاده از داده‌هایی که توسط هاوکینز و همکاران [۵] تولید شده است، عملکرد فاصله استوار در شناسایی نقاط دورافتاده و مقابله با درون‌آوری را زمانی که تعداد مشاهدات زیاد است را مورد بررسی قرار می‌دهیم. سه متغیر اول این مجموعه تشکیل مجموعه داده‌ای با $n = ۷۵$ و $p = ۳$ را می‌دهد که نتایج بررسی در شکل-های ۳ و ۴ آورده شده است.

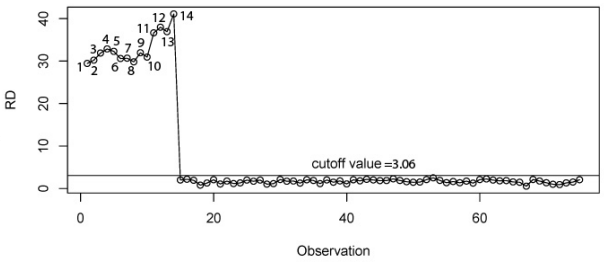
با توجه به شکل ۳ که نشان دهنده فاصله‌های ماهالانوبیس است. با در نظر گرفتن مقدار برشی ۳/۰۶ تنها قادر به شناسایی دو مشاهده ۱۲ و ۱۴ شدیم. زیرا در این روش به علت استفاده از برآوردگرهای کلاسیک در فرمول فاصله ماهالانوبیس با فواصل بزرگ مواجه

نخواهیم شد و بسیاری از مشاهدات دورافتاده به خاطر درون‌آوری شناسایی نمی‌شوند. اما در شکل ۴ چون برآورد




شکل ۳: نمودار فاصله ماهالانوبیس برای داده‌های هاوکینز

بردار میانگین و ماتریس کوواریانس را با روش استوار $MCD$ محاسبه کرده و این برآوردگرها را در فرمول فاصله ماهالانوبیس جایگزین برآوردگرهای کلاسیک نموده‌ایم درون‌آوری از بین رفته و مشاهدات ۱ تا ۱۴ به‌عنوان نقاط دورافتاده شناسایی می‌شوند. همان طور که مشاهده می‌کنیم فاصله این مشاهدات به‌طور معنی‌داری از سایر مشاهدات بیشتر است و مشاهداتی که به‌عنوان مشاهده معمولی شناخته می‌شوند تقریباً فاصله یکسانی با انبوهه داده‌ها دارند.




شکل ۴: نمودار فاصله استوار برای داده‌های هاوکینز

**نکته۲** برآوردگرهای بیضی‌گون کمینه حجم ($MVE$)[۱۰] نیز خواصی مشابه برآوردگرهای $MCD$ دارند و غالباً برآوردگرهای برابری را ارائه می‌دهند. بنابراین فواصل استوار بدست آمده از برآوردگرهای $MVE$ نیز درون‌آوری را از بین می‌برد.

[۱۰] Minimum Volume Ellipsoid

## پیوست

این برنامه برای محاسبه و مقایسه فاصله ماهالانوبیس و فاصله استوار طراحی شده است. ابتدا ماتریس مشاهدات $X$ را به‌عنوان ورودی دریافت کرده و با استفاده از بسته $MASS$ برآوردگرهای $MVE$ و $MCD$ را محاسبه می‌کند. در انتها فاصله ماهالانوبیس و فواصل استوار مبتنی بر برآوردگرهای استوار $MVE$ و $MCD$ را همراه با نمودار این فاصله‌ها ارائه می‌دهد. در این نمودارها فاصله‌ها با مقدار برشی $\sqrt{\chi^2_{۰.۲۵,p}}$ سنجیده شده و نقاطی که بالای خط قرار می‌گیرند، دورافتاده تلقی می‌شوند.

```
RD=function(x){
library(MASS)
n=dim(x)[1];p=dim(x)[2];h=floor((n+p+1)/2)
#================================================
#                compute of MVE &MCD
x.mve=cov.rob(x,h,method="mve")
x.mcd=cov.rob(x,h,method="mcd")
md=numeric();rd.mve=numeric()
rd.mcd=numeric()
#================================================
#    compute of mahalnobis & robust distance
for(i in 1:n){
md[i]=round(sqrt(t(x[i,]-(colMeans(x)))%*%
solve(cov(x)) %*%(x[i,]-(colMeans(x)))),2)
rd.mve[i]=round(sqrt(t(x[i,]-(x.mve$center))%*%
solve(x.mve$cov)%*%(x[i,]-(x.mve$center))),2)
rd.mcd[i]=round(sqrt(t(x[i,]-(x.mcd$center))%*%
solve(x.mcd$cov)%*%(x[i,]-(x.mcd$center))),2)
}
df=data.frame(MD=md,RD_MVE=rd.mve,RD_MCD=rd.mcd)
#================================================
```

```
#     plot of mahalnobis & robust distance
par(mfrow=c(2,1))
plot(md,type='o')
abline((sqrt(qchisq(.975,p))),0)
plot(rd.mve,type='o',lty=2)
abline((sqrt(qchisq(.975,p))),0)
df
}
```

# مراجع


[۱] خلجی، ا. و خردمندنیا، م. (۱۳۹۴). اثر ماسک در شناسایی نقاط دورافتاده چندمتغیره. دهمین سمینار احتمال و فرایندهای تصادفی، ۳۶۵-۳۶۹.

[2] Campbell, N. A. (1989). Bushfire Mapping Using NOAA AVHRR Data. Technical Report, CSIRO.

[3] Donoho, D. L. and Huber, P. J. (1983). The Notion of Breakdown Point. In A Festschrift for Erich L. Lehmann, eds. Bickel, P. J. , Doksum, K. A. and Hodges, J. L. , 157-184.

[4] Hawkins, D. M. (1980). Identification of Outliers, London: Chapman and Hall.

[5] Hawkins, D. M. , Bradu, D. and Kass, G. V. (1984). Location of Several Outlier in Multiple Regression Data Using Elemental Sets. technometrics, 26, 197-208.

[6] Mardia, k., Kent, J. and Bibby, J. (1979). Multivariate Analysis, New York: Academic Press.

[7] Rencher, A. C. (2002). Methods of Multivariate Analysis ($2^{nd}$ED.). New York:Wiley - Interscience.

[8] Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point. In Mathematical Statistics and Application, VOL. B. eds. W. Grossmann, G. Pfug, I. vineze and W. Werz, Dordrecht :Reidel.

[9] Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, 85, 633-639.

[10] Stahel, W. A. (1981). Robuste Schiitzungen: Infinitesimale Optimalitiit und Schiitzungen von Kovarianzmatrizen. Unpublished Ph.D. thesis, ETH Zurich.