برآورد استوار کمینه دترمینان ماتریس کوواریانس

اباذر خلجی، زهرا شاه‌محمدی
گروه آمار، دانشگاه اصفهان

چکیده

نقاط دورافتاده مشاهداتی هستند که از الگوی اکثر داده‌ها پیروی نمی‌کنند. به همین دلیل باعث تخریب برآوردگرهای بردار مکانی و ماتریس پراکندگی می‌شوند. در این مقاله روش کمینه دترمینان ماتریس کوواریانس برای برآورد بردار مکانی و ماتریس پراکندگی معرفی می‌شود. این روش از بین تمام زیرنمونه‌های h نقطه‌ای از n نقطه‌ای، زیر نمونه‌ای را که دارای حداقل دترمینان ماتریس کوواریانس است، اساس محاسبه برآورد بردار مکان و ماتریس پراکندگی قرار می‌دهد. برآوردگرهای بدست آمده از روش کمینه دترمینان ماتریس کوواریانس در مقابل داده‌های دورافتاده مقاوم هستند و علاوه بر این دارای خاصیت هم‌پایائی آفین و بیشینه فروریزش نیز می‌باشند. مشکل عمده این روش زمان‌بر بودن و حجم بالای محاسبات آن می‌باشد، زیرا تعداد زیرنمونه‌های مورد نیاز گاهی اوقات بسیار زیاد می‌شود. برای برطرف نمودن این مشکل الگوریتم‌های سریع معرفی می‌شود که با حداقل باز نمونه‌گیری به نتیجه مطلوب می‌رسند.

واژه‌های کلیدی: برآورد استوار، بیشینه فروریزش، هم‌پایائی آفین.

۱ مقدمه

وجود نقاط دورافتاده در نمونه‌های یک یا چندمتغیره باعث تخریب برآوردگرها اعم از برآوردگر (بردار) مکان و (ماتریس) پراکندگی می‌شود. گاهی اوقات در نمونه‌های یک بعدی از میانه و گاهی نیز از میانگینی که با حذف درصدی از نقاط حدود بالایی و پایینی به دست آمده به عنوان برآورد مکان استفاده می‌شود و سپس با استفاده از همان نمونه تعدیل شده برآورد پراکندگی به دست می‌آید. در حالت تک متغیره امکان شناسایی نقاط دورافتاده از روی نمودار پراکنش داده‌ها نیز وجود دارد. 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.

در شناسایی نقاط دورافتاده زمانی که چند نقطه دورافتاده در مجموع داده‌ها وجود داشته باشد به علت درون‌آوری[۱] شناسایی دیگر امکان‌پذیر نیست. پس مناسب‌تر است از برآوردگرهای استوار[۲] بردار مکانی و ماتریس پراکندگی برای شناسایی نقاط دورافتاده بهره جست. در این مقاله علاوه بر معرفی روش استوار کمینه

[۱] Masking
[۲] Robust

دترمینان ماتریس کوواریانس[۳] (MCD) برای برآورد $\underline{\mu}$ و Σ الگوریتم‌های سریع برای محاسبه این روش نیز ارائه شده است. این روش به دنبال برآورد استواری است که تحت تاثیر نقاط دورافتاده نباشد. این روش نخستین بار توسط روسو [۷] معرفی شده است.

در پاراگراف قبل به استواری یک برآوردگر اشاره شد که در این جا از معیاری ساده و شهودی برای تعریف واژه استوار استفاده می‌کنیم. گوییم یک روش برای برآورد یا آزمون استوار است اگر با حذف یا اصلاح درصد کوچکی از مجموعه داده‌ها نتایج برآورد یا آزمون تغییر قابل توجهی نکند. علاوه بر این‌ها برآوردگرهای MCD دارای خواص زیر نیز هستند:

۱. دارای نقطه فروریزش[۴] ۵۰ درصدی می‌باشند. نقطه فروریزش یک برآوردگر حداقل نسبتی از داده‌ها است که آلوده بودن آنها می‌تواند به برآوردی نامعقول منتهی گردد [۹]،

۲. هم‌پایای آفین[۵] هستند، به این معنی که تغییر مقیاس اندازه گیری بر خصوصیت برآوردگرها اثر نمی‌گذارد [۱۰]،

۳. کارایی بالایی دارند، یعنی واریانس تعمیم یافته یا واریانس کل این برآوردگرها نسبت به سایر برآوردگرها کمتر است. معمولاً رابطه معکوسی بین فروریزش و کارایی بر قرار است به این شکل که افزایش فروریختگی با کاهش کارایی همراه است،

۴. در زمانی معقولی قابل محاسبه هستند، در این جا مقصود از زمانی معقول، کوتاه بودن زمان نیست بلکه بیش از اندازه زیاد و ناممکن نبودن است.

[۳] Minimum Covariance Determinant
[۴] Breakdown Point
[۵] Affine Equivariance

۲ روش کمینه دترمینان ماتریس کوواریانس

تعریف ۱.۲. *اگر n مشاهده p متغیره داشته باشیم، در بین تمام زیرنمونه‌های h نقطه‌ای $(p < h < n)$، بردار میانگین و ماتریس کوواریانس زیرنمونه‌ای که دترمینان ماتریس کوواریانس آن کمینه باشد، برآورد MCD به ترتیب برای $\underline{\mu}$ و Σ می‌باشد [۷].*

کمیت h حجم زیرنمونه انتخابی برای محاسبه برآوردگرهای استوار $\underline{\widetilde{\mu}}$ و $\widetilde{\Sigma}$ می‌باشد و نشان دهنده h مشاهده بالقوه خوب است که در محاسبه برآوردگرهای استوار شرکت دارند. اما $n-h$ مشاهده باقیمانده که باعث افزایش واریانس کل یا واریانس تعمیم‌یافته می‌شوند در محاسبه برآوردگرهای استوار شرکت داده نمی‌شوند و به این ترتیب برآوردگرهای استوار در برابر نقاط دورافتاده مقاوم می‌شوند. داویز [۲] شرح داده است که با قرار دادن $h = \left[\frac{n+p+۱}{۲}\right]$ برآوردگرهای استوار بهبود می‌یابند. لذا ما نیز در محاسبات و مثال‌ها همین مقدار از h را در نظر گرفته‌ایم. در این جا مقصود از علامت $[.]$ جزء صحیح است به طوری که h یک عدد صحیح مثبت است. در عمل h را می‌توان هر عدد صحیح در فاصله $\left[\frac{n+p+۱}{۲}\right] \leq h \leq n$ انتخاب نمود. یکی دیگر از انتخاب‌های مرسوم $h = ۰/۷۵n$ می‌باشد که منجر به افزایش کارایی برآوردگرها در مقابل کاهش نقطه فروریزش آن‌ها می‌شود. اگر $h = n$ آنگاه برآورد MCD مکانی همان بردار میانگین معمولی داده‌ها و برآورد MCD پراکندگی همان برآورد ماتریس کوواریانس نمونه است. در واقع با افزایش h برآوردگرهای

استوار به برآورگرهای معمولی نزدیک می‌شوند که منجر به کاهش نقطه فروریزش می‌شود. تعیین مقدار h بستگی به میزان آلودگی داده‌ها و هدف تحلیل‌گر دارد. بنابراین برای محاسبه برآوردگرهای کمینه دترمینان برای $\underline{\mu}$ و Σ لازم است که تمام N زیرنمونه ممکن تعیین شود که در آن N برابر است با:

$$N = \begin{pmatrix} n \\ h \end{pmatrix} = \frac{n!}{h!\ (n-h)!}.$$

سپس زیر نمونه‌ای که برای آن دترمینان ماتریس کوواریانس کمینه است، برای محاسبه $\widetilde{\mu}$ و $\widetilde{\Sigma}$ مورد استفاده قرار گیرد. واضح است که اگر n بزرگ باشد، تعداد زیرنمونه‌های h عضوی نیز افزایش یافته و به تبعیت از آن حجم محاسبات نیز بسیار زیاد می‌شود. اغلب تلاش‌های دو دهه اخیر مربوط به کم کردن حجم محاسبه‌ها از طریق روش‌های متنوع زیر نمونه گیری تصادفی بوده است.

امروزه به دلایل مختلفی از روش MCD به جای روش بیضی‌گون کمینه حجم[۶] (MVE) استفاده می‌شود که بعضی از این دلایل عبارتند از:

۱. روش MCD دارای کارایی بهتری است چون برآوردگرهای این روش دارای توزیع مجانبی نرمال هستند ولی در مقابل روش MVE دارای نرخ همگرایی پایینی در توزیع است [۷]. برای مثال کارایی مجانبی ماتریس پراکندگی روش MCD با انتخاب $h = ۰/۷۵n$ و $p = ۱۰$ برابر ۴۴ درصد می‌باشد و کارایی ماتریس کوواریانس باز وزن دهی شده با وزن‌هایی که از روش MCD به دست می‌آیند به ۸۳ درصد می‌رسد، در حالی که این مقدار برای روش MVE صفر است.

۲. روسو و فن‌درایسن [۸] نشان دادند که فاصله‌های استواری که بر پایه روش MCD محاسبه می‌شوند بسیار دقیق‌تر از فاصله‌های استوار محاسبه شده بر پایه روش MVE می‌باشند و به همین دلیل بهتر می‌توانند نقاط پرت را شناسایی کنند.

مثال ساده زیر روشن‌گر بسیاری از ایده‌های مربوط به روش MCD می‌باشد.

مثال ۱.۲. *ماتریس داده‌های زیر را در نظر بگیرید.*

$$x = \begin{bmatrix} ۴ & ۱۵ & ۶ & ۱۲ & ۵ \\ ۱۳ & ۲۵ & ۱۲ & ۱۵ & ۱۷ \end{bmatrix}^T$$

در اینجا داریم، $n = ۵$ *و* $p = ۲$*. بنابراین* h *برابر است با:*

$$h = \left[\frac{n+p+۱}{۲}\right] = ۴.$$

تعداد کل زیرنمونه‌های چهارتایی برابر $N = \begin{pmatrix} ۵ \\ ۴ \end{pmatrix} = ۵$ *می‌باشد. این پنج زیر نمونه* ۲×۴ *در زیر آورده شده‌اند:*

$$y_۱ = \begin{bmatrix} ۱۵ & ۲۵ \\ ۶ & ۱۲ \\ ۱۲ & ۱۵ \\ ۵ & ۱۷ \end{bmatrix}, \quad y_۲ = \begin{bmatrix} ۴ & ۱۳ \\ ۶ & ۱۲ \\ ۱۲ & ۱۵ \\ ۵ & ۱۷ \end{bmatrix},$$

[۶] Minimum Volume Ellipsoid

$$y_۳ = \begin{bmatrix} ۴ & ۱۳ \\ ۱۵ & ۲۵ \\ ۱۲ & ۱۵ \\ ۵ & ۱۷ \end{bmatrix}, \quad y_۴ = \begin{bmatrix} ۴ & ۱۳ \\ ۱۵ & ۲۵ \\ ۶ & ۱۲ \\ ۵ & ۱۷ \end{bmatrix},$$

$$y_۵ = \begin{bmatrix} ۴ & ۱۳ \\ ۱۵ & ۲۵ \\ ۶ & ۱۲ \\ ۱۲ & ۱۵ \end{bmatrix},$$

که دترمینان ماتریس کوواریانس این زیرنمونه‌ها برابر است با

$$\det(S_{y_۱}) = ۳۵۶/۳۹, \quad \det(S_{y_۲}) = ۶۱/۵,$$

$$\det(S_{y_۳}) = ۳۶۶, \quad \det(S_{y_۴}) = ۱۵۸/۱۷,$$

$$\det(S_{y_۵}) = ۲۶۲/۳۹,$$

که S_{y_i} ها نشان دهنده ماتریس کوواریانس زیرنمونه i ام می‌باشد. با توجه به این که زیرنمونه $y_۲$ دارای کم‌ترین واریانس تعمیم‌یافته است، همین زیرنمونه را برای محاسبه برآورد MCD برای $\underline{\mu}$ و Σ انتخاب می‌کنیم. بنابراین بردار میانگین و ماتریس کوواریانس زیر نمونه دوم برآوردگر MCD به ترتیب برای $\underline{\mu}$ و Σ می‌باشند. یعنی،

$$\widetilde{\underline{\mu}} = \overline{\underline{y}}_۲ = \begin{bmatrix} ۶/۷۵ \\ ۱۴/۲۵ \end{bmatrix},$$

$$\widetilde{\Sigma} = S_{y_۲} = \begin{bmatrix} ۱۲/۹۱۶۷ & ۱/۴۱۶۷ \\ ۱/۴۱۶۷ & ۴/۹۱۶۷ \end{bmatrix}.$$

در این جا برآورد MCD عیناً برابر برآورد MVE می‌باشد. هر چند در اغلب کاربردها این دو برآورد برابر هم هستند ولی همیشه این دو برآورد یکسان نیستند. اگر الگوریتم محاسبه‌ای فوق را بجای $h = ۴$ برای $h = ۳$ اجرا کنیم زیر ماتریس $\begin{bmatrix} ۴ & ۶ & ۱۲ \\ ۱۳ & ۱۲ & ۱۵ \end{bmatrix}^T$ به‌عنوان زیر ماتریس h نقطه‌ای برای برآورد استوار حاصل می‌شود.

ماتریس کوواریانس داده‌های X تحت تبدیل $A\underline{x} + \underline{b}$ به صورت ASA^T است که در آن S ماتریس کوواریانس داده‌ها می‌باشد. طبق روسو [۷] دترمینان ماتریس کوواریانس داده‌های تبدیل شده برابر است با:

$$det(ASA^T) = det^۲(A) \times det(S),$$

چون $det^۲(A)$ ثابت است، می‌توان نتیجه گرفت برآوردگرهای MCD هم‌پایای آفین هستند. علاوه‌بر این‌ها روسو [۷] نشان داده است که برای هر اندازه نمونه n و هر بعد داده‌های p، نقطه فروریزش برآوردگرهای MCD برابر است با:

$$\frac{\left[\frac{n}{۲}\right] - p + ۱}{n}.$$

اگر n را به سمت بینهایت میل دهیم مقدار فروریزش به ۵۰ درصد همگرا می‌شود که برابر با فروریزش برآوردگرهای MVE می‌باشد.

مثال ۲.۲. *در این مثال به کمک شبیه‌سازی و با استفاده از نرم افزار R عملکرد روش MCD را در آلودگی‌های مختلف مورد بررسی قرار می‌دهیم. ابتدا ماتریس $X_{۶\times ۳}$ را از توزیع $N_۳(\underline{\mu}, \Sigma)$ تولید می‌کنیم. که در آن $\underline{\mu}$ و Σ برابر است با:*

$$\underline{\mu} = \begin{bmatrix} ۱ \\ ۵ \\ ۹ \end{bmatrix}, \quad \Sigma = \begin{bmatrix} ۱ & ۰/۹ & ۰/۹ \\ ۰/۹ & ۱ & ۰/۹ \\ ۰/۹ & ۰/۹ & ۱ \end{bmatrix}.$$

یکی از مشاهدات (مشاهده چهارم) را با مشاهده‌ای که هر بار به روشی آلوده شده، جایگزین می‌کنیم. با توجه به ابعاد ماتریس مشاهدات $h = ۵$ *و تعداد کل زیر نمونه‌های* ۵×۳ *برابر شش می‌باشد. شش زیر نمونه* $y_۱$، $y_۲$، $y_۳$، $y_۴$، $y_۵$ *و* $y_۶$ *به ترتیب از حذف* $\underline{x}_۱^T$، $\underline{x}_۲^T$، $\underline{x}_۳^T$، $\underline{x}_۴^T$، $\underline{x}_۵^T$ *و* $\underline{x}_۶^T$ *از ماتریس مشاهدات* x *بدست می‌آیند که* $\underline{x}_i^T$ *نشان دهنده سطر* i *ام ماتریس* x *می‌باشد. لازم به ذکر است که خروجی برنامه شامل بردار* $\underline{d}$ *و برآوردگرهای استوار مکانی و پراکندگی* MCD *می‌باشد. بردار* $\underline{d}$ *بیانگر مشاهدات انتخاب شده در زیر نمونه‌ای می‌باشد که کمترین دترمینان ماتریس کوواریانس را داراست.*

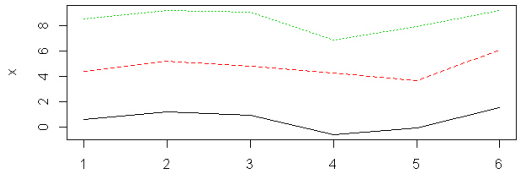
۱. **آلوده کردن میانگین:** در این شبیه‌سازی بردار میانگین متناظر با چهارمین مشاهده را به صورت زیر آلوده می‌کنیم،

$$\underline{\mu}_{outlier} = \underline{\mu} + (۰, ۱, ۰)^T.$$

اما ماتریس کوواریانس را آلوده نمی‌کنیم. به عبارت دیگر برای مشاهده چهارم، میانگین دومین متغیر را به اندازه یک انحراف معیار افزایش می‌دهیم. به این ترتیب ماتریس مشاهدات زیر حاصل می‌گردد.

$$X_۱ = \begin{bmatrix} ۰/۶ & ۴/۳۷ & ۸/۴۸ \\ ۱/۲۳ & ۵/۱۹ & ۹/۱۷ \\ ۰/۹۲ & ۴/۷۶ & ۹/۰۲ \\ -۰/۵۷ & ۴/۲۶ & ۶/۸۲ \\ -۰/۰۳ & ۳/۶۵ & ۷/۸۹ \\ ۱/۵۵ & ۶/۰۷ & ۹/۱۵ \end{bmatrix}.$$

نمودار دنباله‌ای مشاهدات در شکل ۱ ملاحظه می‌شود. با اجرای برنامه آرایه‌های زیر حاصل



شکل ۱: نمودار پراکنش برای آلودگی میانگین

می‌شود:

$$\widetilde{\underline{\mu}} = \begin{bmatrix} ۰/۸۵ \\ ۴/۸۱ \\ ۸/۷۴ \end{bmatrix}, \widetilde{\Sigma} = \begin{bmatrix} ۰/۳۷ & ۰/۵۴ & ۰/۳۲ \\ ۰/۵۴ & ۰/۸۲ & ۰/۴۴ \\ ۰/۳۲ & ۰/۴۴ & ۰/۳۱ \end{bmatrix},$$

$$\underline{d} = \begin{bmatrix} ۱ \\ ۲ \\ ۳ \\ ۵ \\ ۶ \end{bmatrix}.$$

با توجه به بردار $\underline{d}$ زیرنمونه‌ای که مبنای محاسبه برآورد $\underline{\mu}$ و Σ قرار گرفته شامل مشاهده چهارم که آلوده شده بود، نیست. می‌توان نتیجه گرفت روش MCD به‌درستی قادر به شناسایی نقاط پرت می‌باشد زیرا از نمونه‌ای که فاقد

مشاهده آلوده است در برآورد $\underline{\mu}$ و Σ استفاده می‌کند.

۲. **آلوده کردن ماتریس کوواریانس:** در شبیه‌سازی دیگری مشاهدات را از توزیع $N_۳(\underline{\mu}, \Sigma)$ که در بالا شرح داده شد تولید کردیم، اما مشاهده آلوده (مشاهده چهارم) را بر اساس ماتریس کوواریانس زیر تولید نمودیم،

$$\Sigma_{out} = \begin{bmatrix} ۱ & -۰/۹ & ۰/۹ \\ -۰/۹ & ۲ & ۰/۹ \\ ۰/۹ & ۰/۹ & ۱ \end{bmatrix},$$

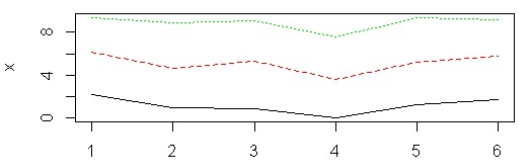
ولی بردار میانگین را تغییر ندادیم. به این ترتیب ماتریس داده‌های زیر حاصل گردید،

$$X_۲ = \begin{bmatrix} ۲/۱۴ & ۶/۱۸ & ۹/۳۲ \\ ۰/۹۸ & ۴/۶۲ & ۸/۹۳ \\ ۰/۸۳ & ۵/۳۲ & ۹/۰۴ \\ ۰/۰۲ & ۳/۵۷ & ۷/۵۰ \\ ۱/۲۶ & ۵/۲۱ & ۹/۳۵ \\ ۱/۷۰ & ۵/۷۵ & ۹/۱۵ \end{bmatrix}.$$

درواقع واریانس متغیر دوم را دو برابر و همبستگی بین متغیر اول و دوم را معکوس نموده‌ایم. اگر داده‌ها را در دو بعد تصور کنیم نمودار پراکنش داده‌ها به شکل بیضی خواهد بود که این نوع آلودگی مشاهده پرت را بر روی خطی موازی محور اصلی بیضی اما با فاصله بیشتری قرار می‌دهد که باعث می‌شود مشاهده آلوده دورتر از سایر مشاهدات قرار بگیرد [۶]. نمودار دنباله‌ای ماتریس داده‌های فوق در شکل ۲ ملاحظه می‌شود.

با اجرای برنامه داریم،

$$\widetilde{\mu} = \begin{bmatrix} ۱/۳۸ \\ ۵/۴۲ \\ ۹/۱۴ \end{bmatrix},$$



شکل ۲: نمودار پراکنش برای آلودگی ماتریس کوواریانس

$$\widetilde{\Sigma}_{(MCD)} = \begin{bmatrix} ۰/۲۹ & ۰/۲۷ & ۰/۰۷ \\ ۰/۲۷ & ۰/۳۵ & ۰/۰۹ \\ ۰/۰۷ & ۰/۰۹ & ۰/۰۵ \end{bmatrix},$$

$$\underline{d} = \begin{bmatrix} ۱ \\ ۲ \\ ۳ \\ ۵ \\ ۶ \end{bmatrix}.$$

با توجه به بردار $\underline{d}$ زیرنمونه‌ای که مبنای محاسبه برآورد $\underline{\mu}$ و Σ قرار گرفته شامل مشاهده چهارم که آلوده شده بود، نیست. می‌توان نتیجه گرفت روش MCD در این نوع آلودگی نیز به‌درستی قادر به شناسایی نقاط پرت می‌باشد زیرا از نمونه‌ای که فاقد مشاهده آلوده است در برآورد $\underline{\mu}$ و Σ استفاده می‌کند. لازم به ذکر است که با ۱۰۰ بار تکرار شبیه سازی، روش MCD در ۹۱ درصد مواقع قادر به ارائه برآوردهای استوار بود. تعداد دفعات باقیمانده ناشی از انحرافات شبیه سازی و خطاهای روش می‌باشد.

۳ الگوریتم پرسرعت محاسبه MCD

برآوردگر MCD از طریق زیر نمونه‌ای که دارای کمترین دترمینان کوواریانس است محاسبه می‌شود و تعداد زیر نمونه‌های ممکن با توجه به مقادیر حجم نمونه n و بعد

متغیر p تعیین می‌شود. از این رو در نمونه‌های با حجم بزرگ اغلب با کثرت زیر نمونه‌ها مواجه هستیم که باعث افزایش مدت زمان مورد نیاز برای محاسبه برآوردگرهای MCD می‌شود. بنابراین ناچار به استفاده از الگوریتم‌هایی هستیم که بتوانیم برآوردگرها را با حداکثر دقت و در حداقل زمان ممکن محاسبه کنیم. تمامی الگوریتم‌هایی که در رابطه با محاسبه برآوردگرهای MCD وجود دارند به نوعی در پی کوتاهترین زمان ممکن از طریق خلاصه سازی در زیر نمونه‌ها برای به‌دست آوردن کمترین دترمینان ماتریس کوواریانس می‌باشند. از این رو روسو و فن‌درایسن [۸] الگوریتم مؤثری را با استفاده از قضیه‌ای که در زیر به آن می‌پردازیم، برای محاسبه برآوردگرهای MCD پیشنهاد کردند و از آن به‌عنوان الگوریتم سریع نام بردند.

قضیه ۱.۳. *فرض کنید* $X_{n\times p} = [\underline{x}_1, \underline{x}_2, ..., \underline{x}_n]^T$، *ماتریس مشاهدات* n *مشاهده* p *متغیره باشد و* H *یک زیر نمونه* h *عضوی تصادفی از بین کلیه زیر نمونه‌های ممکن باشد.* $\underline{\mu}$ *و* S *به ترتیب بردار میانگین و ماتریس کوواریانس این زیر نمونه باشند.*

$$\underline{\mu} = \frac{1}{h}\sum_{i\in H} x_i, \qquad S = \frac{1}{h}\sum_{i\in H}(\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T.$$

اگر $det(S) \neq 0$، *فاصله زیر را تعریف می‌کنیم،*

$$d(i) = \sqrt{(\underline{x}_i - \underline{\mu})^T S^{-1}(\underline{x}_i - \underline{\mu})}, \quad i = 1, 2, ..., n.$$

اکنون H_{new} *زیر نمونه* h *عضوی جدید را طوری انتخاب می‌کنیم که داشته باشیم،*

$$\{d(i);\ i \in H_{new}\} = \{d_{(1)} \leq d_{(2)} \leq ... \leq d_{(h)}\},$$

به‌طوری که $d_{(1)} \leq d_{(2)} \leq ... \leq d_{(n)}$ *فاصله‌های مرتب شده‌اند. اگر* $\underline{\mu}_{new}$ *و* S_{new} *را بر اساس* H_{new} *محاسبه کنیم، آنگاه داریم،*

$$det(S_{new}) \leq det(S).$$

شرط تساوی برقرار است اگر و فقط اگر $\mu = \underline{\mu}_{new}$ *و* $S_{new} = S$ *باشد.*

اثبات. برای اثبات به روسو و فن‌درایسن [۸] مراجعه کنید.

الگوریتم:

اکنون به شرح مراحل اجرای الگوریتم سریع می‌پردازیم:

۱. انتخاب یک زیر نمونه که این انتخاب به دو صورت امکان پذیر است.

- انتخاب یک زیر نمونه h تایی تحت عنوان H_{old} به‌صورت تصادفی تا جایی که $det(S_{old}) > 0$.
- انتخاب یک زیر نمونه $p+1$ تایی y به‌صورت تصادفی و محاسبه $\underline{\mu}_0 = mean(y)$ و $S_0 = Cov(y)$. اگر $det(S_0) = 0$ آنگاه y را توسط اضافه نمودن یک مشاهده تصادفی دیگر بسط داده و آنقدر اضافه نمودن مشاهده را ادامه می‌دهیم تا زمانی که $det(S_0) > 0$ شود. آنگاه فاصله‌های زیر را محاسبه می نماییم،

$$d_0^2 = (\underline{x}_i - \underline{\mu}_0)^T S_0^{-1}(\underline{x}_i - \underline{\mu}_0), \quad i = 1, ..., n,$$

و سپس آنها را به‌صورت صعودی به ترتیب زیر مرتب کرده،

$$(d_۰)_{(۱)} \leq (d_۰)_{(۲)} \leq ... \leq (d_۰)_{(n)},$$

و در نهایت مشاهده‌های متناظر با h فاصله اول را به صورت زیر در زیر نمونه قرار می‌دهیم،

$$H_{old} = \{\underline{x}_{(۱)}, \underline{x}_{(۲)}, ..., \underline{x}_{(h)}\}.$$

۲. محاسبه $\underline{\mu}_{old}$ و S_{old} به همراه کلیه فاصله‌ها، یعنی،

$$d_{old}(i) = \sqrt{(\underline{x}_i - \underline{\mu}_{old})^T S_{old}^{-۱} (\underline{x}_i - \underline{\mu}_{old})}.$$

۳. فاصله‌ها را مرتب کرده و به‌صورت زیر نمایش می‌دهیم،

$$(d_{old})_{(۱)} \leq (d_{old})_{(۲)} \leq ... \leq (d_{old})_{(n)}.$$

۴. مشاهده‌های متناظر با h فاصله اول را در زیر نمونه H_{new} قرار می‌دهیم،

$$H_{new} = \{\underline{x}_{(۱)}, \underline{x}_{(۲)}, ..., \underline{x}_{(h)}\}.$$

۵. $\underline{\mu}_{new}$ و S_{new} را به‌صورت زیر محاسبه می‌کنیم،

$$\underline{\mu}_{new} = mean(H_{new}), \quad S_{new} = Cov(H_{new}).$$

۶. اگر $det(S_{new}) = ۰$ یا $det(S_{old}) = det(S_{new})$ آن‌گاه الگوریتم را متوقف و H_{new} را به‌عنوان بهترین زیر نمونه گزارش می‌کنیم. در غیر این صورت به مرحله دوم باز گشته و با تعویض $\underline{\mu}_{new}$ وS_{new} به ترتیب با $\underline{\mu}_{old}$ و S_{old} الگوریتم را تکرار می‌کنیم.

۷. مرحله دوم تا ششم را با توجه به هزینه‌ها و زمان در دسترس آنقدر ادامه می‌دهیم تا به جواب همگرا برسیم.

همان طور که مشاهده می‌شود این الگوریتم تمامی زیرنمونه‌ها را مورد بررسی قرار نمی‌دهد. بلکه تا آن جایی که ممکن است از تعداد زیرنمونه‌ها می‌کاهد. با توجه به اینکه این الگوریتم فقط با زیرنمونه‌ای که مقدار $\sum_{i=۱}^{h} d(i)$ برای آن کمترین مقدار باشد کار را ادامه می‌دهد، در مقایسه با سایر الگوریتم‌ها از خلاصه سازی و در نتیجه از سرعت بیشتری برخوردار است.

در رابطه با الگوریتم فوق نکات زیر قابل توجه می‌باشند:

- انتخاب عدد کوچك $p + ۲$ به عنوان اندازه نمونه تصادفی با این انگیزه صورت می‌گیرد که احتمال قرار گرفتن یك نقطه دورافتاده در نمونه به حداقل خود برسد [۳].
- منطقی به نظر می‌رسد که تکرار اجرای الگوریتم‌های فوق تا رسیدن به زیرمجموعه‌ای که شامل اکثر مشاهدات بالقوه خوب است ادامه یابد، زیرا این زیرمجموعه دارای کمینه دترمینان ماتریس کوواریانس است و اساس محاسبه برآورد استوار بردار میانگین و ماتریس کواوریانس قرار می‌گیرد. بنابراین برای افزایش دقت الگوریتم‌ها و تضمین رسیدن آن‌ها به یک مینیمم مطلق باید این الگوریتم‌ها در تعداد تکرار بالا انجام شوند، در تکرارهای کم

مینیمم‌های موضعی به عنوان تقریبی مناسب از مینیمم مطلق استفاده می‌گردد.

۴ بررسی داده‌های واقعی

کمپبل [۱] مجموعه داده‌هایی متشکل از $n = ۳۸$ مشاهده در $p = ۵$ بعد را ارائه کرده است. در متون آماری نشان داده شده که ۱۳ مورد از این مشاهدات به شماره‌های ۷ تا ۱۱ و ۳۱ تا ۳۸ نقاط دورافتاده هستند. (در این مورد مراجع [۵] و [۴] را ملاحظه فرمایید.) لازم به ذکر است که روك و ودروف [۱۰] نیز به همین نتایج رسیدند با این تفاوت که مشاهدات ۷، ۱۰، ۱۱ و ۳۱ را مشکوک به دورافتاده بودن و سایر مشاهدات باقیمانده را نقاط دورافتاده شدید[۷] نامیدند. ما از این داده‌ها برای ارائه مثالی عملی از برآوردگر MCD بهره برده‌ایم. نکته بسیار جالب توجه این است که با تعداد تکرار نه چندان زیاد (مثلاً ۵۰ تکرار)، رویت حتی یك مورد از ۱۳ مورد نقطه دورافتاده در باز نمونه نهایی پیشامدی بسیار نادر است. عملاً با افزایش تعداد تکرارها می‌توان احتمال تحقق یافتن این پیشامد را به صفر نزدیك کرد. توجه کنید که از دیدگاه نظری برآوردگرهای MCD یکتا نیستند ولی احتمال یافتن دو زیر مجموعه h عضوی که دترمینان ماتریس کوواریانس آنها برابر کمینه دترمینان ماتریس کوواریانس باشد بسیار ناچیز است. در شکل ۳ در ۴ مورد با ۱۰۰ تکرار بر اساس الگوریتم سریع روسو و فن‌درایسن دروند کاهش دترمینان ماتریس کوواریانس و سرعت همگرایی الگوریتم در روش MCD ملاحظه می‌شود.

در جدول ۱ برآوردهای ML با برآوردهای استوار مقایسه شده‌اند. همان گونه ملاحظه می‌شود در تمام حالت‌های روش استوار برآورد واریانس هر یک از متغیرها از برآورد واریانس روش ML کوچکتر است. واریانس تعمیم یافته که بر اساس روش استوار بدست آمده نیز از واریانس تعمیم یافته بدست آمده از روش ML کمتر است. واریانس متغیرها در جدول ۱ در داخل پرانتز قرار گرفته‌اند.

۵ دستور برنامه R برای محاسبه براوردگرهای MVE و MCD

نرم‌افزار R با استفاده از بسته $(MASS)$ و دستور $cov.rob()$، قادر به محاسبه برآوردگرهای استوار MVE و MCD می‌باشد. دستور به‌صورت زیر می باشد،

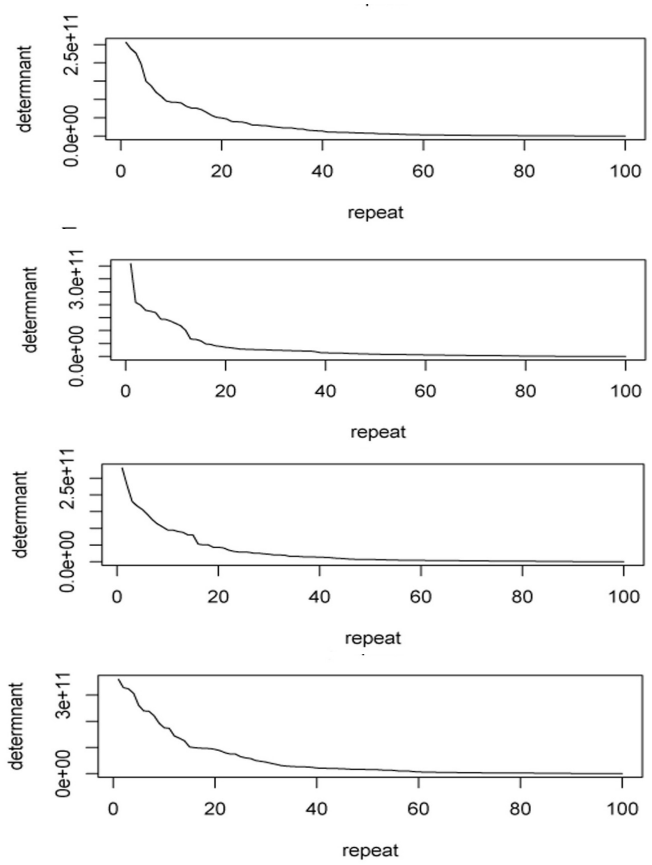
```
cov.rob(x, cor = FALSE,
 quantile.used = floor((n + p + 1)/2),
        method = c("mve", "mcd")),
```

که در آن x ماتریس داده‌ها، عملگر منطقی cor با اختیار دو مقدار $TRUE$ و $FALSE$ در خروجی برنامه به ترتیب اجازه نمایش و عدم نمایش ماتریس همبستگی را می‌دهد. با استفاده از کد $quantile.used$ می‌توان حجم زیرنمونه مورد استفاده را مشخص کرد و با استفاده از کد $method$ یکی از دو روش MCD یا MVE انتخاب می‌شود. برای جزئیات بیشتر می‌توان به $help$ برنامه R مراجعه کرد.

۶ نتیجه گیری

با توجه به مباحثی که عنوان شد اگر مجموعه داده‌ها آلوده باشد، برآوردگرهای کلاسیک تحت تاثیر نقاط دور افتاده

[۷] Extrem Outlier



شکل ۳: روند کاهش حجم بر اساس الگوریتم سریع روسو و فن‌درایسن در ۱۰۰ تکرار

قرار گرفته و کارایی لازم را ندارند. اگر آلودگی ناشی از خطاهای اندازه گیری و ثبتی باشد می‌توان پس از حذف این نقاط از برآوردگرهای کلاسیک بهره جست اما به دلیل اینکه شناسایی نقاط دور افتاده چندمتغیره بسیار سخت‌تر از حالت تک‌متغیره است این امر نیز امکان ناپذیر است. بنابراین نیازمند روش‌هایی هستیم که در برابر نقاط دور افتاده برآوردگرهای معقولی را ارائه دهند. از این رو روشی که در این مقاله مورد بحث قرار گرفته حتی زمانی که مشکوک به وجود نقاط دورافتاده در مجموعه داده‌ها هستیم مفید بوده و برآوردگرها (بردار میانگین و ماتریس پراکندگی) را بر اساس داده‌هایی که به هم نزدیکتر هستند محاسبه می‌کند. یعنی اگر مجموعه داده‌ها در فضای چند بعدی به صورت یک توده تصور شود در برآورد از نمونه‌ای که در مرکز توده است و دارای کمترین ماتریس پراکندگی می‌باشد محاسبه می‌شود.

مراجع

[1] Campbell, N. A. (1989). Bushfire mapping using NOAA AVHRR data. Technical Report, CSIRO.

[2] Davies, P. L. (1987). Asymptotic bi-

جدول ۱: مقایسه برآوردهای ML و استوار

برآورد	روش		ML		
میانگین	۲۸۶/۶	۲۲۷/۸	۲۸۸/۶	۱۲۹/۱	۱۰۳/۶
ماتریس کوواریانس	−۵۱۶	−۶۳۹	−۲۰۹۱	۵۶۵	(۴۰۶)
	−۹۴۳	−۱۱۸۴	−۳۲۵۸	(۱۲۲۵)	۵۶۵
	۹۰۲۱	۱۱۰۶۰	(۳۱۴۰۵)	−۳۲۵۸	−۲۰۹۱
	۳۳۴۰	(۴۱۰۳)	۱۱۰۶۰	−۱۱۴۸	−۶۳۹
	(۲۷۲۲)	۳۳۴۰	۹۰۲۱	−۹۴۳	−۵۱۶
واریانس تعمیم یافته	۱۱۹۵۲۰۰۰۰۰۰۰				

برآورد	روش		استوار		
میانگین	۲۷۷/۱	۲۱۵/۳	۲۶۲/۲	۱۴۹/۶	۱۰۹/۴
ماتریس کوواریانس	−۲۸۹	−۱۶۶	۲۰۰۰	۴۸۱	(۳۸۶)
	−۳۴۶	−۱۹۱	۲۴۶۶	(۶۰۴)	۴۸۱
	−۱۶۶۹	−۱۰۱۲	(۹۷۶۷)	۲۴۶۶	۲۰۰۰
	۲۸۰	(۲۳۶)	−۱۰۱۲	−۱۹۱	−۱۶۶
	(۳۷۷)	۲۸۰	−۱۶۶۹	−۳۴۶	−۲۸۹
واریانس تعمیم یافته	۱۲۳۶۹/۶۷				

(1995). The Behavior of the stahel-donoho robust multivariate location and shape. Journal of the American Statistical Association, 4, 51-67.

[6] Rencher, A. C. (2002). Methods of multivariate analysis (2^{nd} ED.). New York: Wiley - interscience.

[7] 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.

[8] Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the min-

havior of s-stimators of multivariate location parameteres and dispersion matrices. The Annals of Statistics, 15, 1269-1292.

[3] Hawkins, D. M. and Olive, D. J. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and new algorithm (with discussion). Journal of the American Statistical Association, 97, 136-159.

[4] Kosinski, A. S. (1999). A procedure for the detection of multivariate outlier. Computational Statistic and Data Analysis, 29, 145-161.

[5] Maronna, R. A. and Yohai, V. J.

imum covariance determinant estimator. Technometrics, 41(3), 212-223.

[9] Sakata, S. and With, H. (1998). Breakdown point. Encyclopedia of Statistical Sciences, 2, 84-89.

[10] Woodruff, D. L, and Rocke, D. M. (1996). Identification of outlier in multivariate data. Journal of the American Statiatical Association, 91(435), 1047-1061.