مجلهی بررسیهای آمار رسمی ایران
سال۲۲، شمارهی۱، بهار و تابستان ۱۳۹۰، صص ۳۳ - ۴۴

مروری بر براوردگر ماتریس کوواریانس با کم‌ترین دترمینان و کاربرد آن

فاطمه سادات حسینی بهارانچی[†,*] و مسعود یارمحمدی[‡]

[†] مؤسسه‌ی آموزش عالی ایوانکی
[‡] دانشگاه پیام نور مرکز تهران

چکیده: هدف اصلی این مقاله معرفی روشی جهت شناسایی نقاط دورافتاده در مجموعه داده‌های چندمتغیره است. روش استوار به‌کار گرفته شده در این مقاله روش ماتریس کوواریانس با کم‌ترین دترمینان[۱] (MCD) است. به علاوه به دو ویژگی مهم براوردگرهای استوار یعنی نقطه فروریزش و تابع نفوذ اشاره می‌کنیم. سپس به معرفی عامل سازگاری و عامل تصحیح نمونه‌ی متناهی در براوردگر MCD خواهیم پرداخت. در پایان با ارایه‌ی یک مثال کاربردی کارایی روش MCD را با روش کلاسیک در رابطه با شناسایی داده‌های دورافتاده بررسی می نماییم.

واژگان‌کلیدی: براوردگر ماتریس کوواریانس با کم‌ترین دترمینان، شناسایی دورافتاده‌ها، فواصل استوار.

۱- مقدمه

همواره آمارشناسان علاقمند به یافتن مشاهدات دورافتاده بوده‌اند. یک مشاهده‌ی دورافتاده داده‌ای است که به‌طور آشکار از دیگر اعضای داده‌ها فاصله دارد. داده‌هایی که به درستی وارد نشده‌اند و یا از جامعه‌ای متفاوت از جامعه داده‌ها می‌آیند، موجب اریبی براوردها و نتایج گمراه‌کننده می‌شوند. داده‌های دورافتاده با مخدوش کردن مفروضات برخی از مدل‌های آماری نتایج نادرستی را به همراه دارند. نه تنها در مجموعه داده‌های یک‌متغیره بلکه در مجموعه داده‌های چندمتغیره نیز بررسی، شناسایی و کاهش اثرات

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۸۹/۵/۱۷، پذیرش: ۱۳۹۰/۱/۳۰.

مشاهدات دورافتاده از اهمیت بسیار زیادی برخوردار است. بـه‌طور کلـی تجزیـه و تحلیـل داده‌های چندمتغیره نیازمند به کارگیری روش‌های آماری است که تحت تأثیر آلـودگی‌هـای[۲] موجود در مشاهدات قـرار نگیـرد و یـا بـه‌عبـارت دیگـر در مواجهـه بـا داده‌هـای دورافتـاده استوار باشند. از روش‌های استوار برای شناسایی داده‌های دورافتـاده در مـدل‌هـای آمـاری چنـدمتغیره مـی‌تـوان بـه براوردگرهـای M، براوردگرهـای S، براوردگـر بـا کـم‌تـرین حجـم بیضیگون[۳] (MVE) و براوردگر ماتریس کوواریانس با کم‌ترین دترمینـان (MCD) اشـاره نمود.

مسئله‌ایی که نباید آن را از نظر دور داشت این است که روش‌های استوار زیادی ارایـه شده است، اما تنها برخی از آن‌ها در شناسایی و لحاظ کردن مشاهدات دورافتاده به‌خـوبی عمل می‌کنند. روش‌های کلاسیک معرفی‌شده برای شناسایی مشاهدات دورافتاده به‌راحتـی قابـل محاسـبه بـوده لـیکن قابـل اعتمـاد نیـستند. در مقابـل، روش‌هـای اسـتوار مـستلزم محاسبات پیچیده بوده اما نتایج دقیق‌تری بـه همـراه دارنـد و آمارشناسـان نیـز همـواره بـه دنبال روش قابل اطمینانی برای شناسایی نقاط دورافتاده در مجموعه داده‌های چنـدمتغیره هستند. در این مقاله یکـی از روش‌هـای اسـتوار بـه نـام مـاتریس کوواریـانس بـا کـم‌تـرین دترمینان را مورد بحث قرار می‌دهیم. این روش استوار علاوه بر ویژگی‌های مطلوب آماری، از نظر شهودی نیز به‌راحتـی قابـل درک اسـت. در بخـش دوم ایـن مقالـه نگـاهی بـه روش کلاسیک شناسایی مشاهدات دورافتاده در مجموعـه داده‌هـای چنـدمتغیره خـواهیم داشـت. سپس به تعریف روش ماتریس کوواریـانس بـا کـم‌تـرین دترمینـان و الگـوریتم محاسـبه آن می‌پردازیم. در ادامـه‌ی همـین بخـش دو ویژگـی مهـم براوردگرهـای اسـتوار بـه نـام نقطـه‌ی فروریزش و تابع نفـوذ را تعریـف مـی‌کنـیم. در بخـش سـوم مقالـه دربـاره عامـل سـازگاری MCD و نیز عامل تصحیح نمونه‌ی متناهی براوردگر MCD اشـاراتی خـواهیم داشـت. در بخش چهارم این مقاله به ارایه‌ی یک مثال کاربردی می‌پردازیم.

۲- شناسایی دورافتاده‌ها

نقاط دورافتاده موجـود در داده‌هـای یـک‌بعـدی و یـا دوبعـدی از روی نمودارهـای سـاده‌ای مانند نمودار جعبه‌ای و نمودار پراکنش به‌راحتـی قابـل تـشخیص هـستند. لـیکن شناسـایی آن‌ها در ابعاد بالاتر از دو، از روی نمودار مشکل است. علت این امر آن است که داده‌های دورافتاده در میان داده‌های چندمتغیره مخفی شده و به‌راحتـی نمایـان نمـی‌شـوند. تـا کنـون

روش‌های مختلفی برای شناسایی دورافتاده‌ها معرفی شده است [۵ و ۹]. روش کلاسیک شناسایی دورافتاده‌ها محاسبه‌ی فاصله هر نقطه از مرکز داده‌ها با در نظر گرفتن شکل داده‌ها است. بدین صورت که یک دورافتاده، نقطه‌ایی است که فاصله‌ی آن نقطه از یک نقطه برش از پیش تعیین‌شده بزرگ‌تر باشد. این معیار کلاسیک که فاصله‌ی ماهالانوبیس[۴] نام دارد، از یک نقطه‌ی $x \in R^p$ با بردار مکان $T \in R^p$ با ماتریس مقیاس معلوم C (ماتریس متقارن مربعی و معین مثبت با بعد p) به‌صورت زیر تعریف می‌شود:

$$d_C(x,T) = \sqrt{(x-T)^t C^{-1}(x-T)}. \qquad (۱)$$

در این رابطه، بردار میانگین ($\bar{x}$) و ماتریس کوواریانس (S) به دست آمده از نمونه را به‌ترتیب به‌عنوان بردار مکان (T) و ماتریس مقیاس (C) که در (۱) آمده است به کار می‌گیریم. هرگاه مجذور فاصله‌ی ماهالانوبیس مشاهده، بالای چندک $\alpha-$۱ی توزیع خی‌دو با p درجه‌ی آزادی قرار گیرد آن مشاهده به‌عنوان دورافتاده در نظر گرفته می‌شود. انتخاب نقطه‌ی برش مذکور از این رو است که اگر $X \sim N_p(\mu,\Sigma)$ باشد آنگاه فواصل $d^2_\Sigma(X,\mu)$ دارای توزیع χ^2_p است. با فرض اینکه $\bar{x}$ و S مقادیری نزدیک به μ و Σ بگیرند اغلب فواصل $d^2_S(X,\bar{x})$ نیز با $d^2_\Sigma(X,\mu)$ هم‌توزیع هستند.

روسو و ون‌زمرن [۷] پیشنهاد کردند با قرار دادن براوردهای استوار به جای T و C در رابطه‌ی (۱) فواصل استوار ساخته شود. با این کار یک روش استوار جهت شناسایی دورافتاده‌ها حاصل می‌شود. در این مقاله روش ماتریس کوواریانس با کم‌ترین دترمینان (MCD) معرفی شده که توسط روسو [۶] ارایه شده است. علت انتخاب این روش این است که علی‌رغم نزدیکی به بردار میانگین و ماتریس کوواریانس کلاسیک، نقاط دورافتاده را در محاسبات لحاظ نمی‌کند. به‌طور دقیق‌تر تعریف این براوردگر به‌صورت زیر است:

اولین گام برای یافتن براوردگر MCD از نمونه‌ایی به حجم n، شامل یافتن زیرنمونه‌ایی به حجم h $(\frac{n}{2} \le h \le n)$ از n مشاهده است به‌طوری که واریانس تعمیم‌یافته‌ی (دترمینان ماتریس واریانس-کوواریانس) این زیرنمونه کمینه گردد. سپس بردار میانگین و ماتریس کوواریانس زیرنمونه‌ی بهینه‌ی به دست آمده به‌ترتیب براوردگر مکان MCD و براوردگر مقیاس MCD هستند. از دیدگاه استنباط آماری براوردگرهای

MCD ویژگی‌های خوب تئوری از جمله سازگاری و به‌طور مجانبی کارا بودن [۱] را دارا هستند.

الگوریتم FAST-MCD یکی از الگوریتم‌های محاسبه‌ی براوردگر MCD است که توسط روسو و وندریزن [۹] ارایه شده و روش کارامدی است که هم اکنون برای محاسبه‌ی این براوردگر موجود می‌باشد. در زیر به اختصار به C-STEP که برای اجرای الگوریتم FAST-MCD لازم است، اشاره می‌شود.

C-STEP: مجموعه داده‌های $X_n=\{x_1,\ldots,x_n\}$ از مشاهدات p بعدی را در نظر گرفته و فرض کنید $H_1\subset\{1,\ldots,n\}$ با اندازه‌ی $|H_1|=h$ و T_1 و S_1 به‌ترتیب بردار میانگین حسابی و ماتریس کوواریانس h مشاهده باشند.

$$T_1=\frac{1}{h}\sum_{i\in H_1}x_i$$

$$S_i=\frac{1}{h}\sum_{i\in H_1}(x_i-T_1)(x_i-T_1)^t$$

اگر $\det(S_1)\neq 0$، آنگاه فواصل نسبی به‌صورت

$$d_i(i)=\sqrt{(x_i-T_1)^t S_1^{-1}(x_i-T_1)} \qquad i=1,2,\ldots,n$$

تعریف می‌شوند. بعد از مرتب کردن فواصل نسبی به‌صورت ترتیبی $(d_1)_{1:n}\leq(d_1)_{2:n}\leq\ldots\leq(d_1)_{n:n}$، H_2 با حجم h به‌گونه‌ای انتخاب می‌شود که: $\{d_1(i),i\in H_2\}=\{(d_1)_{1:n},\ldots,(d_1)_{h:n}\}$. حال S_2 و T_2 مربوط به H_2 را محاسبه کرده و داریم $\det(S_2)\leq\det(S_1)$. حالت تساوی برقرار می‌شود، اگر و تنها اگر $S_1=S_2$ و $T_1=T_2$ (اثبات [۹]).

در C-STEP فرض شده است که $\det(S_1)\neq 0$. اگر $\det(S_1)=0$ آنگاه به تقریب MCD با کم‌ترین دترمینان دست یافته‌ایم و اگر $\det(S_1)>0$ با به کارگیری C-STEP، S_2 به گونه‌ای تعیین می‌شود که $\det(S_2)\leq\det(S_1)$. در این روش روی h مشاهده با کم‌ترین فواصل متمرکز می‌شویم که S_2 از S_1 متمرکزتر است. با تکرار C-STEP یک روند تکراری حاصل می‌شود: اگر $\det(S_2)=0$ یا $\det(S_2)\geq\det(S_1)$، توقف

می‌کنیم. در غیر این صورت با به کارگیری C-STEP، $\det(S_3)$ را به دست می‌آوریم و الی آخر. دنباله‌ی $\det(S_1) \geq \det(S_2) \geq \det(S_3) \geq \ldots$ نامنفی است و بنا بر این همگرا نیز هست. در عمل معمولاً کم‌تر از ۱۰ تکرار برای رسیدن به نتیجه‌ی مطلوب کافی است و بعد از آن، اجرای مجدد C-STEP مقدار دترمینان را کاهش نمی‌دهد. ایده‌ی اجرای الگوریتم FAST-MCD این است که تعداد زیادی زیرمجموعه H_1 انتخاب و C-STEP را برای هر یک تا مرحله‌ی همگرایی ادامه داده می‌شود. سپس از بین جواب‌ها، جوابی برگزیده می‌شود که از بقیه دترمینان کم‌تری دارد. چگونگی انتخاب زیرمجموعه‌های ابتدایی H_1 روش‌های مختلفی دارد [۹].

۱-۲- نقطه‌ی فروریزش (Breakdown Point)

یکی از ویژگی‌های مهم براوردگرهای استوار از جمله براوردگر MCD نقطه‌ی فروریزش است. میزان عدم حساسیت یک براوردگر استوار نسبت به نقاط دورافتاده نقطه‌ی فروریزش نامیده می‌شود. به‌عبارت دیگر، نقطه‌ی فروریزش کوچک‌ترین کسری از آلودگی‌هاست که سبب می‌شود براوردگر مقداری دورتر از مقدار واقعی خود بگیرد. از کل n مشاهده تنها h مشاهده در براورد MCD به کار رفته و $n-h$ مشاهده در براورد MCD نقشی ندارند. بنا بر این طبق تعریف، نقطه‌ی فروریزش تقریباً برابر با $(n-h)/n$ خواهد بود. براوردگر MCD بیش‌ترین مقدار فروریزش خود را زمانی که $h=[(n+p+1)/2]$ باشد خواهد داشت [۹]. طبق تعریف نقطه‌ی فروریزش عبارت است از:

$$\varepsilon_n^*(C_n, Z_n) = \min_{1\leq m\leq n}\left\{\frac{m}{n} : \sup_{Z'_n} D\left(C_n(Z_n), C_n(Z'_n)\right) = \infty\right\}$$

که در آن $D(A,B) = \max\left\{\left|\lambda_1(A) - \lambda_1(B)\right|, \left|\lambda_p(A)^{-1} - \lambda_p(B)^{-1}\right|\right\}$ و $\lambda_p(A) \leq \ldots \leq \lambda_1(A)$ مقادیر ویژه‌ی ماتریس A هستند. می‌توان نشان داد نقطه‌ی فروریزش میانگین نمونه $\frac{1}{n}$ و نقطه‌ی فروریزش میانه ۵۰٪ است [۲]. در اینجا لازم به ذکر است که نقطه‌ی فروریزش براوردگر MCD به ۵۰٪ هم می‌رسد.

۲-۲- تابع نفوذ (Influence Function)

از دیگر مفاهیم مهم در بحث براوردگرهای استوار، مفهوم تابع نفوذ است. تابع نفوذ بیانگر میزان اریبی است که با اضافه کردن کسر ناخالصی‌ها به میزان ε بر روی براورد حاصل می‌شود. براوردگری که نسبت به نقاط دورافتاده استوار است، دارای تابع نفوذ کراندار می‌باشد. یعنی براوردگر استوار، براوردگری است که اجازه نمی‌دهد تغییرات نمونه باعث ایجاد تغییرات بزرگ در آن شود [۹]. طبق تعریف تابع نفوذ عبارت است از:

$$IF(z,T,H)=\lim_{\varepsilon\to 0}\frac{T\left(\left(1-\varepsilon\right)H+\varepsilon\Delta_z\right)-T\left(H\right)}{\varepsilon}=\frac{\partial}{\partial\varepsilon}T\left((1-\varepsilon)H+\varepsilon\Delta_z\right)$$

می‌توان نشان داد میانگین نمونه استوار نیست زیرا تابع نفوذ آن با توجه به تغییرات کراندار نمی‌باشد [۲]. در اینجا لازم به ذکر است که براوردگر MCD دارای تابع نفوذ کراندار است.

۳- بهبود براوردگر MCD

خاصیت سازگاری یک معیار مجانبی است که رفتار براوردگر را وقتی که اندازه‌ی نمونه بزرگ باشد ارزیابی می‌کند. برای آنکه براوردگر مقیاس MCD به‌طور مجانبی با مدل مفروض برای مشاهدات سازگار گردد در عامل سازگاری ضرب می‌شود. برای نمونه‌های کوچک می‌توان براورد کوواریانس MCD را در عامل تصحیح نمونه‌ی متناهی نیز ضرب کرد تا اریبی آن تصحیح شود. در ادامه‌ی این بخش درباره‌ی این دو عامل مطالبی را مطرح می‌کنیم.

۱-۳- عامل سازگاری[۵]

براوردگر اولیه‌ی مکان MCD به‌صورت زیر تعریف می‌شود:

$$\hat{\mu}_n^{raw}=\frac{1}{h}\sum_{j=1}^{h}x_{i_j}$$

و براورد اولیه‌ی کوواریانس MCD عبارت است از:

$$\hat{\Sigma}_n^{raw} = \frac{c}{h}\sum_{j=1}^{h}\left(x_{i_j} - \hat{\mu}_n^{raw}\right)\left(x_{i_j} - \hat{\mu}_n^{raw}\right)^t$$

که در آن c عامل سازگاری است. وارد کردن ثابت c به محاسبات به این علت است که معمولاً می‌خواهیم براوردگرها با مدلی که از پیش برای جامعه تعیین کرده‌ایم به‌طور مجانبی سازگار باشند. باتلر و همکاران [۱] نشان دادند که هرگاه توزیع مورد نظر متعلق به خانواده‌ی بیضوی با چگالی

$$f_{\mu,\Sigma}(x) = \frac{1}{\sqrt{\det(\Sigma)}}\, g\left[(x-\mu)^t\,\Sigma^{-1}(x-\mu)\right]$$

(تابع $g: R^+ \to R^+$ یک تابع غیرصعودی است که تضمین می‌کند یک توزیع تک‌مدی است) باشد آنگاه cیی وجود دارد به‌طوری که

$$.\; c\,\hat{\Sigma}_n^{raw} \xrightarrow{P} \Sigma$$

که در آن ثابت c به‌صورت زیر تعریف می‌شود:

$$c_1 = \frac{\dfrac{h}{n}}{P\left(\chi^2_{p+2} < \chi^2_{p,\frac{h}{n}}\right)},$$

که در آن χ^2_υ یک متغیر تصادفی خی‌دو با υ درجه‌ی آزادی است و $\chi^2_{\upsilon,\varepsilon}$ نقطه‌ی برش متغیر تصادفی χ^2_υ است. به‌علاوه روسو و وندریزن [۹] در الگوریتم FAST-MCD، عامل سازگاری دیگری که در اینجا آن را c_2 می‌نامیم پیشنهاد کردند:

$$c_2 = \frac{med\left\{d^2_{\hat{\Sigma}_n}\left(x_1,\hat{\mu}_n\right),\ldots,d^2_{\hat{\Sigma}_n}\left(x_n,\hat{\mu}_n\right)\right\}}{\chi^2_{p,0/5}}$$

که در آن $\hat{\mu}_n$ و $\hat{\Sigma}_n$ بردار میانگین و ماتریس کوواریانس نمونه‌ی MCD است. لازم به ذکر است که در همه‌ی نرم‌افزارهایی که براوردگر MCD را محاسبه می‌کنند عامل سازگاری اعمال می‌گردد.

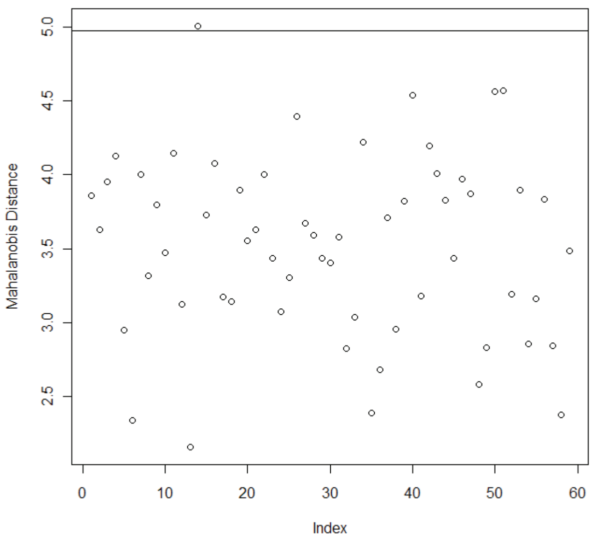
۲-۳- عامل تصحیح نمونه‌ی متناهی[۶]

پیزن و همکاران [۵] توانستند به کمک شبیه‌سازی ثابتی به‌عنوان عامل تصحیح نمونه‌ی متناهی ارایه کنند تا در نمونه‌های کوچک اریبی براورد ماتریس کوواریانس MCD تصحیح گردد. لیکن اشکال این عوامل تصحیح این است که مانند یک جعبه‌ی سیاه عمل می‌کند. به این معنا که این عوامل خود توسط شبیه‌سازی‌ها و روابط درون‌یابی به دست آمده و هم اکنون در پیش فرض نرم‌افزارها قرار گرفته‌اند و هیچ رابطه‌ی نظری دقیقی برای آن‌ها وجود ندارد.

۴- ارایه‌ی یک مثال

در این بخش از مقاله به شناسایی مشاهدات دورافتاده در مجموعه داده‌ای شامل ۵۹ مشاهده (نوشیدنی‌های مختلف) و ۱۳ متغیر (ترکیبات آن‌ها) می‌پردازیم [۵].

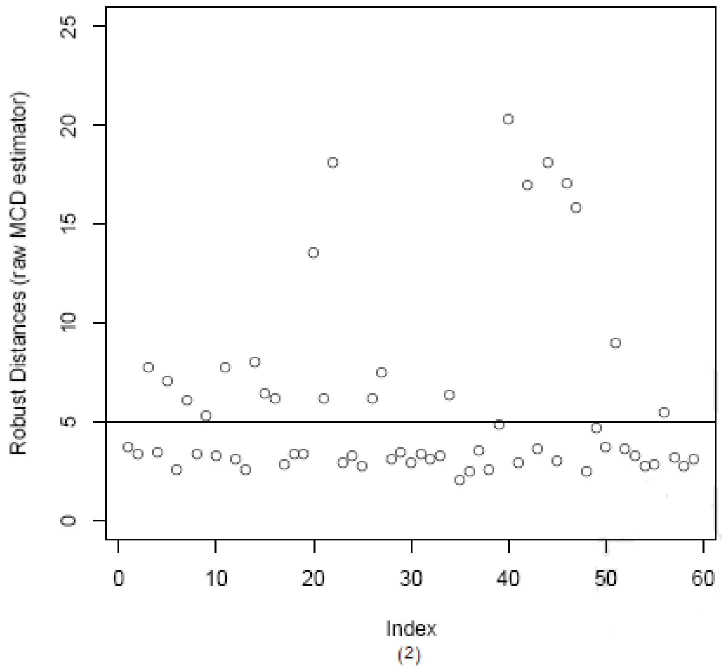
در شکل ۱ فواصل کلاسیک ماهالانوبیس برای این مجموعه داده رسم شده است. در این شکل به جز یک مشاهده بالای نقطه‌ی برش که در اینجا $\sqrt{\chi^{۲}_{۱۳,۰/۹۷۵}}$ انتخاب شده است، نقطه‌ی دورافتاده دیگری دیده نمی‌شود. در شکل ۲ براوردهای



شکل ۱- فواصل ماهالانوبیس برای مجموعه داده‌ی نوشیدنی‌ها

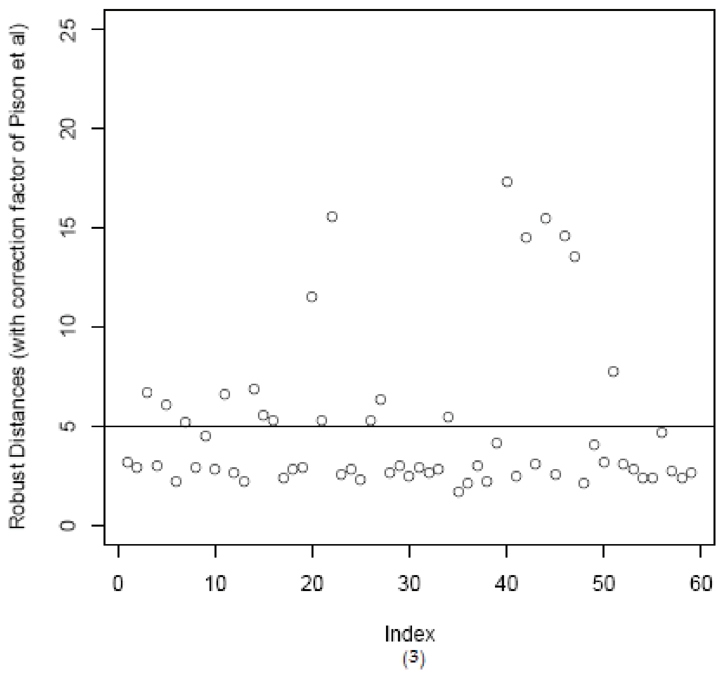
اولیه‌ی MCD (که عامل سازگاری در آن اعمال شده) رسم شده است. این شکل ۱۷ مشاهده‌ی دورافتاده (بالای نقطه‌ی برش $\sqrt{\chi^{2}_{13,0/975}}$) را نشان می‌دهد. به‌طور ویژه ۷ نقطه از آن‌ها فاصله‌ی بیش‌تری نسبت به نقطه‌ی برش دارند.

در شکل ۳ فواصل استوار با استفاده از براوردهای MCD (که عامل سازگاری در آن اعمال شده) پس از اعمال عامل تصحیح نمونه‌ی متناهی رسم شده است. این نمودار ۱۹ مشاهده‌ی دورافتاده (بالای نقطه‌ی برش $\sqrt{\chi^{2}_{13,0/975}}$) را نشان می‌دهد. البته با توجه به شکل‌های ۲ و ۳ دیده می‌شود که قبل و پس از اعمال عامل تصحیح نمونه‌ی متناهی تنها ۷ نقطه از مشاهدات قرارگرفته بالای نقطه‌ی برش فواصل استوار

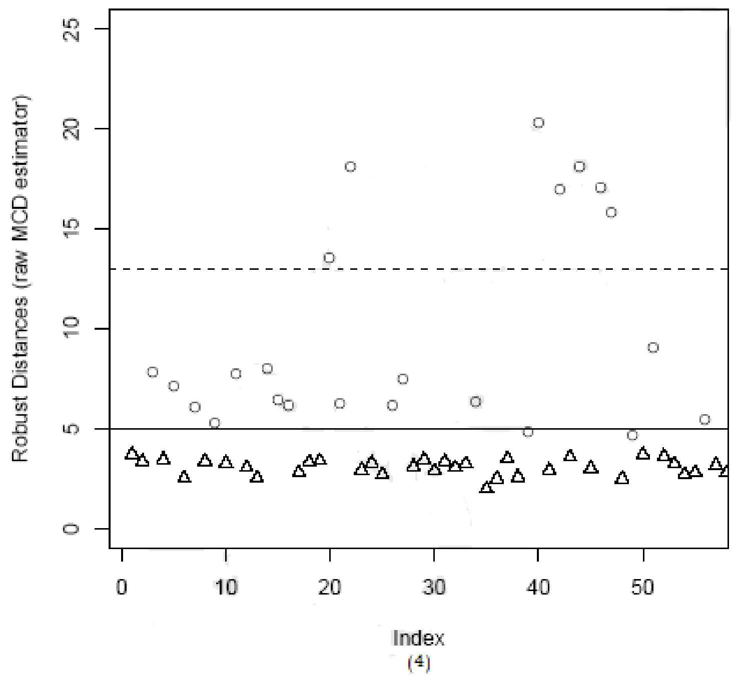


شکل ۲- فواصل استوار MCD اولیه برای مجموعه داده‌ی نوشیدنی‌ها

بزرگی داشته و به‌طور آشکاری دورافتاده هستند. لذا برای بررسی دقیق‌تر دیگر نقاط می‌توان از تعدیل ارایه‌شده توسط هاردین و راک [۳] استفاده کرد. هاردین و راک معتقدند بهتر است در روش معمول MCD به‌جای استفاده از نقطه‌ی برش بر اساس توزیع خی‌دو نقاط با چندکی از یک توزیع F مقایسه شوند. با توجه به این‌که بیان این روش خارج از حوصله‌ی این مقاله است برای جزئیات بیش‌تر به [۳] مراجعه شود. شکل ۴ کاربرد این تعدیل برای داده‌های مثال نوشیدنی‌ها را نشان می‌دهد. نقطه‌ی برش در این روش تغییر کرده و تعداد نقاط دورافتاده به ۷ مورد تقلیل می‌یابد که در مقایسه با شکل‌های ۲ و ۳ به نظر می‌رسد نتایج دقیق‌تری را نشان می‌دهد.



شکل ۳- فواصل استوار MCD برای مجموعه داده‌ی نوشیدنی‌ها پس از اعمال عامل تصحیح نمونه‌ی متناهی



شکل ۴- فواصل استوار MCD برای مجموعه داده‌ی نوشیدنی‌ها پس از به کارگیری تعدیل هاردین و راک

توضیحات

1. Minimum Covariance Determinant
2. Contamination
3. Minimum Volume Ellipsoid
4. Mahalanobis Distance
5. Consistency Factor
6. Finite-sample Correction Factor

مرجع‌ها

[1] Butler, R.W., Davies, P.L. and Jhun, M. (1993). Asymptotics for the Minimum Covariance Determinant Estimator. *The Annals of Statistics*, **21**, 1385-1400.

[2] Filzmoser, P., Maronna, R. and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, **52**, 1694–1711.

[3] Hardin, J. and Rocke, D. (2005). The Distribution of Robust Distances. *Journal of Computational and Graphical Statistics*. **14**, 928–946.

[4] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons, New York.

[5] Pison, G., Van Aelst, S. and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, **55**, 111–123.

[6] Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point. In: Mathematical Statistics and Applications, Vol. B, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: Reidel, 1985, 283–297.

[7] Rousseeuw, P.J. and Van Zomeren, B.C. (1991). Robust distances: simulations and cutoff values. In: Directions in Robust Statistics and Diagnostics 2. W. Stahel and S. Weisberg, New York, 195–203.

[8] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.

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

فاطمه سادات حسینی بهارانچی
فوق لیسانس آمار
شهر ایوانکی، بلوار آیت‌ا... طالقانی، خیابان دانشگاه، مؤسسه‌ی آموزش عالی ایوانکی.
رایانشانی: hosseini_153@yahoo.com

مسعود یارمحمدی
دکتری آمار
تهران، دانشگاه پیام نور مرکز تهران، سازمان مرکزی دانشگاه پیام نور، گروه آمار.
رایانشانی: masyar@pnu.ac.ir