مجله علوم آماری، پاییز و زمستان ۱۴۰۰
جلد ۱۵، شماره ۲، ص ۴۴۳ – ۴۶۲
DOI: 10.29252/jss.15.2.443
مقاله پژوهشی

شناسائی نقاط دورافتاده چندمتغیره بر اساس تابع دورافتادگی ژرفا-مبنا

سکینه دهقان، محمدرضا فریدروحانی

گروه آمار، دانشکده علوم ریاضی، دانشگاه شهید بهشتی

تاریخ دریافت: ۱۳۹۹/۱۲/۰۸ تاریخ پذیرش و انتشار: ۱۴۰۰/۰۲/۰۴

چکیده: تابع ژرفا با در نظرگرفتن ویژگی‌های هندسی مجموعه داده‌های چندمتغیره و رتبه‌بندی مشاهدات ابزار مناسبی را در آمار ناپارامتری چندمتغیره فراهم آورده است. به عبارت دیگر، این تابع منجر به مرتب‌سازی از مرکز به بیرون نقاط چندمتغیره می‌شود. از آن‌جا که دورافتادگی نقاط به طور اجتناب‌ناپذیری وابسته به ترتیب داده‌ها است، این مرتب‌سازی می‌تواند راهی برای شناسایی نقاط دورافتاده فراهم کند. در این مقاله، بر اساس مفهوم تابع ژرفا، یک روش ناوردای آفین برای شناسائی نقاط دورافتاده چندمتغیره بیان می‌شود. ویژگی مطلوب ناوردای آفین تضمین می‌کند که نقطه دورافتاده تحت هرگونه تبدیل از محورهای مختصات کماکان به‌عنوان دورافتاده شناسایی شود. پیاده‌سازی این روش نسبت به بیشتر روش‌های چندمتغیره که دارای پیچیدگی محاسباتی هستند، ساده‌تر است. بر اساس مطالعات شبیه‌سازی عملکرد روش پیشنهادی بر اساس توابع ژرفای مختلف مورد بررسی قرار گرفته است. سرانجام، روش بیان شده برای داده‌های مسکن شهرهای منتخب ایران در سال ۱۳۹۷، بکار برده می‌شود.

واژه‌های کلیدی: تابع ژرفا، تابع دورافتادگی، دورافتاده، ناوردایی آفین.

آدرس الکترونیکی نویسنده مسئول مقاله: سکینه دهقان، Sa_dehghan@sbu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62G99, 62H99.

۱ مقدمه

در بسیاری از مسائل استنباطی، شناسایی نقاط دورافتادهٔ[1] نمونه، به دو دلیل از اهمیت فراوانی برخوردار است. اولاً حضور نقاط دورافتاده در یک مجموعه داده اثر نامطلوبی بر استنباط از آن مجموعه داده گزارده و منجر به برآوردگرها، آماره‌های آزمون، بازه‌های اطمینان و ... نامناسبی می‌شود. ثانیاً برخی موارد نقاط دورافتاده اطلاعات مفیدی از مدل داده‌ها به تحلیل‌گر می‌دهند و از این‌روشناسائی آن‌ها برای تحلیل‌گر دارای اهمیت است. شناسایی نقاط دورافتاده در پزشکی، کشف تقلب، یافتن خطای اندازه‌گیری، رکوردگیری‌های ورزشی و ... کاربرد دارد. برای شناسایی نقاط دورافتاده در حالت یک‌متغیره، روش‌های متعددی وجود دارد (هاوکینز ، ۱۹۸۰؛ بارنت و لویس ، ۱۹۹۴). دورافتاده در نظر گرفتن نقاط خارج از بازه‌ای با مرکز میانگین مشاهدات و کران‌هایی که با اضافه و کم کردن سه برابر انحراف معیار مشاهدات از میانگین حاصل می‌شوند، یکی از متداولترین روش‌ها برای شناسائی نقاط دورافتاده یک‌متغیره است. این روش برگرفته از ایدهٔ بازهٔ اطمینان با ضریب اطمینان ۹۹٫۸۷ برای میانگین توزیع نرمال است. البته برخی از نویسندگان با توجه به نوع مساله از ۲٫۵ یا ۲ برابر انحراف استاندارد حول میانگین استفاده می‌کنند (میلر ، ۱۹۹۱). میانگین و انحراف معیار مشاهدات برآوردگرهای استواری از پارامترهای مکان و مقیاس نیستند و عملکرد این روش که معمولا تحت برقراری فرض نرمال از آن استفاده می‌شود، تحت تاثیر نقاط دورافتاده است. یک روش جایگزین مناسب، استفاده از بازه‌ای بر اساس برآوردگر استوار میانه مشاهدات برای پارامتر مکان و برآوردگر استوار میانه قدرمطلق انحراف نقاط از میانه برای پارامتر مقیاس است. بسط روش‌های یک‌متغیره به چندمتغیره با مشکلات زیادی همراه است. با به کارگیری روش‌های یک‌متغیره و بررسی دورافتادگی حاشیه‌ای نقاط چندمتغیره، می‌توان به طور مجزا در هر بعدی نسبت به شناسائی نقاط دورافتاده اقدام کرد. روش بررسی دورافتادگی حاشیه‌ای نمی‌تواند بسیاری از نقاط دورافتاده را شناسایی کند و همچنین ویژگی‌های هندسی مجموعهٔ داده‌ها را در نظر نمی‌گیرد. تجسم مجموعهٔ داده‌ها برای ابعاد بیشتر از سه بسیار سخت است، بنابراین روش تصویری‌سازی عملاً محدود به بعد سه و کمتر است و برای ابعاد بالاتر کاربردی ندارد. همچنین می‌توان با روش‌های خوشه‌بندی چندمتغیره، داده‌ها را به خوشه‌هایی تقسیم کرد و خوشه‌هایی که فقط شامل یک یا دو نقطه هستند، به صورت دورافتاده در نظر گرفت. به دلیل سختی خوشه‌بندی، زمان‌بر بودن محاسبات رایانه‌ای، وابستگی به بعد داده‌ها، اندازه نمونه و ... این روش خیلی به‌کار نمی‌رود. شناسائی نقاط دورافتاده چندمتغیره بر اساس فاصله ماهالانوبیس توسط راک و وودراف (۱۹۹۶) معرفی شد. با اینکه این روش به طور گسترده‌ای به کار می‌رود، ولی چنانچه

[1] Outlier

فرض تقارن بیضوی برای مشاهدات برقرار نباشد، دارای عملکرد مناسبی نیست. به منظور افزایش کارائی این روش در حالتی که فرض تقارن بیضوی برقرار نیست، سرفلینگ و موزوندر (۲۰۱۳) با به کارگیری روش جستجوی تصویر، شناساگر دورافتادگی معرفی شده توسط راک و وودراف (۱۹۹۶) را تعمیم دادند. کان و همکاران (۲۰۱۵) با بهکارگیری چندین فاصله، ابزارهای تشخیصی متعددی را برای شناسائی نقاط دورافتاده چندمتغیره معرفی کردند.

با معرفی تابع ژرفای نیم‌فضا[1] توسط توکی (۱۹۷۵) در بسیاری از مباحث آمار ناپارامتری از این تابع استفاده می‌شود. تابع ژرفا[2] با در نظر گرفتن ویژگی‌های هندسی مجموعه داده‌های چندمتغیره و رتبه‌بندی مشاهدات چندمتغیره ابزار مناسبی را فراهم آورده است. به عبارت دیگر، در حالت چندمتغیره، ژرفای داده‌ها می‌تواند برای اندازه‌گیری ژرفا یا دورافتادگی یک نقطهٔ چندمتغیره نسبت به توزیع آن به‌کار رود که قابلیت مزبور منجر به مرتب‌سازی مرکز به بیرون نقاط نمونه می‌شود. واضح است، این مرتب‌سازی می‌تواند راهی برای شناسایی نقاط دورافتاده فراهم کند و نقطه دورافتاده تحت هرگونه تبدیل از محورهای مختصات کماکان به‌عنوان دورافتاده شناسایی شود. با توجه به ویژگی‌های مطلوب تابع ژرفا، پژوهش‌هائی برای شناسائی نقاط دورافتاده انجام گرفته است. چن و همکاران (۲۰۰۸) تابع ژرفای فضائی هسته‌ای[3] را معرفی کردند و بر اساس آن یک شناساگر برای تعیین نقاط دورافتاده تعیین کردند. آن‌ها نشان دادند که این تابع ویژگی‌های هندسی توزیع را کاملا در نظر گرفته و دارای عملکرد مناسبی در تشخیص نقاط دورافتاده است. البته روش معرفی شده توسط چن و همکاران (۲۰۰۸) دارای پیچیدگی‌های محاسباتی در تعیین آستانهٔ مورد نیاز برای شناسایی نقاط دورافتاده است. دنگ و سرفلینگ (۲۰۱۰) با بهکارگیری توابع ژرفا، شناساگرهایی با نقطه فروریزش[4] بالا معرفی کردند. نقطه فروریزش که عبارت است از نسبتی از نمونه که می‌تواند آلوده شود بدون این‌که بعضی نقاط دورافتاده مخفی شوند (به صورت غیر دورافتاده شناسایی شوند)، معیاری برای بررسی استواری [5] شناساگرهای دورافتادگی محسوب می‌شود (روزبه و امینی ، ۱۳۹۸؛ روزبه و معنوی ، ۱۳۹۹). در ادامه، دوودو و چاکرابوتی (۲۰۱۳) شناساگرهای دنگ و سرفلینگ را برای حالت خاص توزیع چوله نرمال به‌کار بردند. فن (۲۰۱۶) با محاسبه ژرفای نقاط اصلی نمونه نسبت به مجموعه نقاط اصلی نمونه و نقاطی که نسبت به ژرفترین نقطه متقارن شده‌اند، چنین استدلال کرد که می‌توان عملکرد روش شناسائی نقاط دورافتاده بر اساس تابع ژرفا را بهبود بخشید. روش وی با در نظر گرفتن نمونه متقارن شده، دارای پیچیدگی

[1] Half space depth
[2] Depth function
[3] Kernelized spatial depth function
[4] Breakdown point
[5] Robustness

محاسباتی است و نسبت به شناساگرهای دنگ و سرفلینگ (۲۰۱۰) باعث بهبود عملکرد چشم‌گیری در زمینه شناسائی نقاط دورافتاده نشده است.

در بخش ۲ مفهوم ژرفای داده و انواع تابع ژرفا مرور می‌شوند، سپس تابع دورافتادگی ژرفا–مبنا معرفی می‌شود. در بخش ۳ شناساگرهای دورافتادگی[1] بر اساس توابع ژرفا بر اساس یک مدل آلودگی معرفی شده و به طور دقیق مقدار آستانه مورد نیاز برای شناسائی نقاط دورافتاده مشخص می‌شود. مطالعات شبیه‌سازی برای بررسی و مقایسه عملکرد توابع ژرفای مختلف در بخش ۴ ارائه می‌شود. در بخش ۵ بر اساس داده‌های قیمت مسکن شهرهای منتخب ایران که توسط مرکز آمار ایران ثبت شده‌اند، شهرهایی که از نظر ارزش مالی مسکن نسبت به سایر شهرها دورافتاده محسوب می‌شوند، شناسائی و نتایج تحلیل می‌شوند.

۲ توابع ژرفا و دورافتادگی

فرض کنید اندازهٔ احتمال P بر فضای $\mathcal{R}^p$ داده شده و F تابع توزیع متناظر با آن باشد. ژرفای داده، اندازه‌ای از میزان ژرف بودن یا دورافتاده بودن یک نقطه نسبت به تودهٔ داده‌ها یا توزیع F است. هر تابع $D(\mathbf{x}, F)$ که ترتیبی از مرکز به بیرون از نقاط $\mathbf{x}$ متعلق به $\mathcal{R}^p$ فراهم کند، تابع ژرفای متناظر با F نامیده می‌شود. بنابراین بر اساس تابع ژرفای مفروض، نقاط p متغیره به نحوی رتبه‌بندی می‌شوند که ژرفای بزرگتر نشان‌دهندهٔ مرکزی‌تر بودن نقاط و ژرفای کوچکتر نشان‌دهندهٔ دورافتاده‌تر بودن آن‌ها است. تحت تابع ژرفای مفروض، نقطه یا میانگین نقاط با بیشترین مقدار ژرفا، مرکز را تشکیل می‌دهند.

تعریف ۱. (ژو و سرفلینگ ، ۲۰۰۰) اگر $\boldsymbol{X}$ یک بردار تصادفی p بعدی بر فضای احتمال $(\Omega, \mathcal{F}, P)$ باشد، تابع $D(\cdot, F) : \mathcal{R}^p \longrightarrow \mathcal{R}$، ژرفای آماری است، اگر در ویژگی‌های زیر صدق کند:

الف– ناوردایی آفین: $D(\boldsymbol{x}, F)$ مستقل از دستگاه مختصات باشد. یعنی برای هر بردار تصادفی $\boldsymbol{x}$ در R^p، هر ماتریس $\boldsymbol{A}$ی نامنفرد $p \times p$ و هر بردار p بعدی $\boldsymbol{b}$، $D\left(\boldsymbol{A}\, \boldsymbol{x} + \boldsymbol{b}, F_{\boldsymbol{AX}+\boldsymbol{b}}\right) = D\left(\boldsymbol{x}, F_{\boldsymbol{X}}\right)$، که در آن $F_{\boldsymbol{X}}$ و $F_{\boldsymbol{AX}+\boldsymbol{b}}$ به‌ترتیب توابع توزیع بردارهای تصادفی $\boldsymbol{X}$ و $\boldsymbol{A}\, \boldsymbol{X} + \boldsymbol{b}$ هستند.

ب– ماکسیمم شدن در مرکز: اگر F نسبت به $\boldsymbol{\theta}$ متقارن باشد، آنگاه $D(\boldsymbol{x}, F)$ در $\boldsymbol{\theta}$ ماکسیمم است.

ج– به صفر رسیدن در بینهایت: اگر $\|\boldsymbol{x}\| \to \infty$ آنگاه $D\left(\boldsymbol{x}, F\right) \to ۰$.

د– یکنوائی نسبت به ژرفترین نقطه: اگر برای هر بردار تصادفی $\boldsymbol{x}$ در $\mathcal{R}^p$، $D\left(\boldsymbol{\theta}, F\right) \geq D\left(\boldsymbol{x}, F\right)$، آنگاه برای هر $\alpha \in [۰, ۱]$، $D\left(\boldsymbol{\theta} + \alpha\left(\boldsymbol{x} - \boldsymbol{\theta}\right), F\right) \geq D\left(\boldsymbol{x}, F\right)$.

[1] Outlier identifiers

به منظور تصویر و مرتب‌سازی داده‌های چندمتغیره توکی (۱۹۷۵) تابع ژرفای نیم‌فضا، (HD)، را معرفی کرد. به دنبال آن تابع ژرفای سادکی[۲] (LD)؛ (لیو ، ۱۹۸۸)،، تابع ژرفای ماهالانوبیس[۳] (MD)؛ (لیو و همکاران ، ۱۹۹۳)، تابع ژرفای تصویر-پایه[۱] (PD)؛ (ژو ، ۲۰۰۳)، و تابع ژرفای فضائی[۲] (SD)؛ (وردی و ژانگ ، ۲۰۰۰)، پیشنهاد شد، که در ادامه معرفی می‌شوند.

تعریف ۲. (توکی ، ۱۹۷۵) تابع ژرفای نیم‌فضا برای $\boldsymbol{x} \in \mathcal{R}^p$ نسبت به توزیع F به صورت

$$HD(\boldsymbol{x}, F) = \inf \left\{ P(H) : \text{ نیم‌فضایی بسته شامل } \boldsymbol{x} \text{ در فضای } p \text{ بعدی است } H \right\}$$

تعریف می‌شود، که در آن $P(H)$ احتمال نیم‌فضای H توسط توزیع F است. ژرفای نیم‌فضای نقطهٔ $\boldsymbol{x}$ نسبت به نمونهٔ $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ در $\mathcal{R}^p$ عبارت است از کمترین کسر نمونه در بین تمامی نیم‌فضاهای بستهٔ شامل نقطهٔ $\boldsymbol{x}$. به عبارت دیگر

$$HD(\boldsymbol{x}, F_n) = \frac{1}{n} \min_{\|\boldsymbol{u}\|=1} \# \left\{ i : \boldsymbol{u}^T \boldsymbol{X}_i \leq \boldsymbol{u}^T \boldsymbol{x}, i = 1, \ldots, n \right\}$$

که در آن F_n تابع توزیع تجربی نمونهٔ $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ است.

تعریف ۳. بردارهای $k_1, \ldots, k_t$ در R^p به طور آفین[۳] مستقل هستند، اگر و تنها اگر $t-1$ بردار $k_2 - k_1, \ldots, k_t - k_1$ به طور خطی مستقل باشند. در فضای اقلیدسی p بعدی R^p، مجموعه‌ای از نقاط که ترکیبی محدب از $p+1$ نقطه به طور آفین مستقل باشند را سادک می‌نامند. به عبارت دیگر اگر نقاط به طور آفین مستقل را با $k_1, \ldots, k_{p+1}$ نشان داده شوند، سادک تولید شده توسط این نقاط به صورت

$$S(k_1, k_2 \ldots, k_{p+1}) = \left\{ k : k = a_1 k_1 + \ldots + a_{p+1} k_{p+1}, a_i \geqslant 0, \sum_{i=1}^{p+1} a_i = 1 \right\}.$$

تعریف می‌شود، که در آن $k_1, \ldots, k_{p+1}$ رئوس سادک نام دارند.

---

[2] Simplicial depth
[3] Mahalanobis depth
[1] Projection-based depth
[2] Spatial depth
[3] Affine invariance

تعریف ۴. تابع ژرفای سادکی (لیو ، ۱۹۸۸) برای $\boldsymbol{x} \in \mathcal{R}^p$ نسبت به توزیع F به صورت

$$LD\left(\boldsymbol{x}, F\right) = P\left\{\boldsymbol{x} \in S\left[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{p+1}\right]\right\},$$

تعریف می‌شود، که در آن $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{p+1}$ مشاهدات مستقل از توزیع F هستند و $S\left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p+1}\right]$ نشان‌دهندهٔ سادک بسته در $\mathcal{R}^p$ با رئوس $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p+1}$ است. ژرفای سادکی نقطهٔ $\boldsymbol{x}$ نسبت به نمونهٔ $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ در $\mathcal{R}^p$ برابر با

$$LD\left(\boldsymbol{x}, F_n\right) = \binom{n}{p+1}^{-1} \sum_{*} I\left(\boldsymbol{x} \in S\left[\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_{p+1}}\right]\right)$$

است، که در آن $*$ شامل همهٔ زیرمجموعه‌های $p+1$ عضوی از $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ و $I(\cdot)$ تابع نشانگر است.

تعریف ۵. (لیو و همکاران ، ۱۹۹۳) تابع ژرفای ماهالانوبیس برای $\boldsymbol{x} \in R^p$ نسبت به توزیع F

$$MD\left(\boldsymbol{x}, F\right) = \left[1 + \left(\boldsymbol{x} - \boldsymbol{\mu}_F\right)^T \sum_{F}^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}_F\right)\right]^{-1}$$

است، که در آن $\boldsymbol{\mu}_F$ و $\boldsymbol{\Sigma}_F$ به‌ترتیب بردار میانگین و ماتریس کوواریانس توزیع F هستند. برای به دست آوردن نسخهٔ نمونه‌ای ژرفای ماهالانوبیس نقطهٔ $\boldsymbol{x} \in R^p$، کافی است به جای $\boldsymbol{\mu}_F$ و $\boldsymbol{\Sigma}_F$ برآوردهای نمونه‌ای آن‌ها جایگزین شوند، که یک انتخاب معمول به ترتیب بردار میانگین نمونه $\bar{\boldsymbol{X}}$ و ماتریس کوواریانس نمونه‌ای $\boldsymbol{S}$ است.

تعریف ۶. فرض کنید $\boldsymbol{X}$ دارای توزیع F باشد و

$$\tilde{O}_P\left(\boldsymbol{x}, F\right) = \sup_{\|\boldsymbol{u}\|=1} \left| \frac{\boldsymbol{u}^T\boldsymbol{x} - \mu\left(F_{\boldsymbol{u}^T\boldsymbol{X}}\right)}{\sigma\left(F_{\boldsymbol{u}^T\boldsymbol{X}}\right)} \right| \qquad (1)$$

که در آن $\mu(\cdot)$ و $\sigma(\cdot)$ به ترتیب اندازهٔ یک متغیرهٔ مکان و مقیاس توزیع F هستند. هرگاه هر دو مقدار صورت و مخرج کسر (۱) صفر یاشند، یعنی $\boldsymbol{u}^T\boldsymbol{x} - \boldsymbol{\mu}\left(F_{\boldsymbol{u}^T\boldsymbol{X}}\right) = \sigma\left(F_{\boldsymbol{u}^T\boldsymbol{X}}\right) = 0$، آنگاه $\left|\frac{u^T\boldsymbol{x} - \mu\left(F_{\boldsymbol{u}^T\boldsymbol{X}}\right)}{\sigma\left(F_{\boldsymbol{u}^T\boldsymbol{X}}\right)}\right|$ صفر در نظر گرفته می‌شود. ژرفای تصویر-مبنا (ژو ، ۲۰۰۳) برای $\boldsymbol{x} \in \mathcal{R}^p$ نسبت به

توزیع F به صورت $PD\left(\boldsymbol{x},F\right)=\left[1+\tilde{O}_P\left(\boldsymbol{x},F\right)\right]^{-1}$، تعریف می‌شود.

برای به دست آوردن نسخهٔ نمونه‌ای ژرفای تصویر-مبنا به جای توزیع F همتای نمونه‌ای آن، F_n جایگزین می‌شود. هر زوج خاصی که برای برآورد (μ,σ) به‌کار رود، تابع ژرفای نمونه‌ای تصویر-مبنا خاصی را نتیجه می‌دهد. اما به نظر انتخابی مناسب می‌تواند زوج $(\mu,\sigma)=(Med,MAD)$ باشد، که در آن Med میانه و MAD میانهٔ قدرمطلق انحراف از میانه هستند. برای نمونهٔ یک‌متغیرهٔ $\boldsymbol{Y}_N=\{Y_1,\ldots Y_N\}$ در R، به صورت زیر تعریف می‌شوند.

$$Med\left(\boldsymbol{Y}_N\right)=\text{median}\left\{Y_i,1\leq i\leq N\right\}, \tag{۲}$$

$$MAD\left(\boldsymbol{Y}_N\right)=\text{median}\left\{\left|Y_i-Med\left(\boldsymbol{Y}_N\right)\right|,1\leq i\leq n\right\}. \tag{۳}$$

تعریف ۷. تابع علامت برداری در $\mathcal{R}^p$ به صورت

$$\boldsymbol{S}(\boldsymbol{x})=\begin{cases}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \boldsymbol{x}\neq 0\\ 0 & \boldsymbol{x}=0\end{cases}$$

تعریف می‌شود، که در آن $||\boldsymbol{x}||$ نرم اقلیدسی $\boldsymbol{x}$ به صورت $\sqrt{\boldsymbol{x}^T\boldsymbol{x}}$ است. اگر $\boldsymbol{X}$ دارای توزیع F باشد، آن‌گاه ژرفای فضایی (وردی و ژانگ ، ۲۰۰۰) نقطهٔ $\boldsymbol{x}$ نسبت به توزیع F به صورت

$$SD\left(\boldsymbol{x},F\right)=1-\left\|E\left(\boldsymbol{S}(\boldsymbol{x}-\boldsymbol{X})\right)\right\|$$

تعریف می‌شود. مقدار تابع فضایی نمونه‌ای نقطهٔ $\boldsymbol{x}$، نسبت به نمونهٔ $\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\}$ نیز عبارتست از

$$SD\left(\boldsymbol{x},F_n\right)=1-\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{S}(\boldsymbol{x}-\boldsymbol{X}_i)\right\|.$$

با توجه به تعریف توابع ژرفا واضح است، نقطه‌ای با مقدار ژرفای کمتر به عنوان دورافتاده تلقی می‌شود. بر این اساس، دنگ و سرفلینگ (۲۰۱۰)، برای به دست آوردن تابع دورافتادگی ابتدا تابع ژرفا را بر بازهٔ ۰ تا ۱ استاندارد کرده، سپس از رابطهٔ $O(\mathbf{x},F)=1-D(\mathbf{x},F)$ تابع دورافتادگی متناظر را تعریف کردند. به طور مشابه، دهقان و فریدروحانی (۲۰۱۹) برای تابع ژرفای استاندارد شده بر بازهٔ ۰

تا ۱، تابع دورافتادگی را به صورت $O(\mathbf{x}, F) = \frac{۱}{D(\mathbf{x}, F)} - ۱$ تعریف کردند، که مقادیر خود را در بازه $[۰, \infty)$ اختیار می‌کند. در فضای $\mathcal{R}^p$، منحنی‌های تراز متناظر تابع $g(\mathbf{x})$ عبارتند از رده‌هایی هم‌ارزی از نقاط $\mathbf{x}$ که برای آن‌ها مقدار $g(\mathbf{x})$ یکسان است. به طور مثال اگر $g(.)$ تابع چگالی احتمال باشد، منحنی‌های تراز، مقادیر چگالی احتمال روی نقاط واقع در ردهٔ هم‌ارزی را مشخص می‌کنند. این رده‌های هم‌ارزی لزوماً با رده‌های دورافتادگی یکسان نیستند. چنان‌چه $g(.)$ را تابع ژرفا یا تابع دورافتادگی در نظر بگیریم، رده‌های هم‌ارزی متناظر، شامل نقاطی با مقدار ژرفای برابر خواهند بود. منحنی‌های تراز در تحلیل چندمتغیره نوعاً به دو دلیل توصیف چگالی احتمال و شناسایی نقاط دورافتاده به‌کار می‌روند. مجموعهٔ

$$\{\mathbf{x} \in \mathcal{R}^p : \quad O(\mathbf{x}, F) = t\} \tag{۴}$$

مجموعهٔ سطح یا منحنی تراز دورافتادگی t نامیده می‌شود.

۳ کشف نقاط دورافتاده چندمتغیره

برای تعیین این‌که نقطهٔ $\mathbf{x} \in \mathcal{R}^p$ نسبت به تابع توزیع F دورافتاده است یا خیر، ابتدا تابع دورافتادگی $O(., F)$ و مقدار آستانه‌ای λ مشخص می‌شود. سپس نقاطی که مقدار دورافتادگی آن‌ها از این مقدار آستانه‌ای تجاوز کند، نقاط دورافتاده نسبت به توزیع F نامیده می‌شوند. به عبارت دیگر، هر نقطهٔ $\mathbf{x}$ متعلق به $\text{out}(\lambda, F) = \{\mathbf{x} : O(\mathbf{x}, F) > \lambda\}$، ناحیهٔ λ-دورافتاده از توزیع F نامیده می‌شود. حال باید مقدار آستانهٔ λ تعیین شود. از طرف دیگر، در عمل، توزیع F معمولاً نامعلوم است و باید از تابع دورافتادگی نمونه‌ای استفاده کرد. ابتدا مدل آلودگی $F = (۱ - \varepsilon) G + \varepsilon H$ را برای توزیع F در نظر می‌گیریم، که G توزیع ایده‌آل معلوم مدل، H منبع نامعلوم آلودگی و ε نسبت و یا احتمال دورافتادگی است. مشاهدات کرانگین نسبت به توزیع G به اشتباه به عنوان دورافتاده و نسبت به توزیع H به درستی به عنوان دورافتاده تلقی می‌شوند. از این‌رو مشاهدات کرانگین نسبت به توزیع G مثبت نادرست و نسبت به توزیع H مثبت درست، نامیده می‌شوند. نرخ مثبت نادرست [۱] را حاصل‌ضرب احتمال این‌که $\mathbf{X}$ از توزیع G تولید شده باشد در احتمال این‌که تحت توزیع G به عنوان λ-دورافتاده شناسایی شود، تعریف می‌کنیم. بنابراین نرخ مثبت نادرست به صورت $(۱ - \varepsilon) P_G(O(\mathbf{X}, G) > \lambda)$، تعریف می‌شود و به

[1] False positive rate

طور مشابه نرخ مثبت درست[۲] $\varepsilon P_H(O(\mathbf{X},H)>\lambda)$، است. آستانهٔ λ باید به گونه‌ای انتخاب شود که برای λ به اندازهٔ کافی بزرگ، $P_G(O(\mathbf{X},G)>\lambda)$ مقداری کوچک و برای λ به اندازهٔ کافی کوچک، $P_H(O(\mathbf{X},H)>\lambda)$ مقداری بزرگ باشد. بنابراین می‌توان به طور تقریبی روابط

$$(۱-\varepsilon)\,P_G(O(\mathbf{X},G)>\lambda)\cong P_G(O(\mathbf{X},G)>\lambda),$$
$$\varepsilon P_H(O(\mathbf{X},H)>\lambda)\cong\varepsilon,$$

را در نظر گرفت. اگر $\alpha = P_G(O(\mathbf{X},G)>\lambda)$، در این صورت به طور تقریبی α و ε به ترتیب نشان‌دهندهٔ نرخ مثبت نادرست و نرخ مثبت درست هستند. اگر α از قبل تعیین شده باشد، آنگاه

$$\lambda = F^{-۱}_{O(\mathbf{X},G)}(۱-\alpha). \qquad (۵)$$

بنابراین آستانهٔ λ چندک $۱-\alpha$ام توزیع $O(\mathbf{X},G)$ تحت توزیع ایده‌آل G است. مطلوب آن است که نرخ مثبت نادرست نسبت به نرخ مثبت درست مقداری کوچک باشد، بنابراین اگر $\delta = \alpha/\varepsilon$، آنگاه δ باید مقداری کوچک باشد. در عمل δ و ε از قبل تعیین شده هستند، و لذا رابطهٔ (۵) به صورت

$$\lambda = F^{-۱}_{O(\mathbf{X},G)}(۱-\delta\varepsilon) \qquad (۶)$$

است. برای مثال اگر $\delta = ۰/۱$ و نسبت دورافتادگی $\varepsilon = ۰/۲۵$ باشد، آنگاه $\lambda = F^{-۱}_{O(\mathbf{X},G)}(۰/۹۷۵)$. منطقی است، نسبت آلوده شده از نمونه با افزایش اندازه نمونه کاهش یابد. در این صورت، می‌توان ε را برابر $\frac{c}{\sqrt{n}}$ که c مقداری ثابت است، تعریف کرد. بنابراین از رابطهٔ (۶) داریم

$$\lambda_n = F^{-۱}_{O(\mathbf{X},G)}\left(۱-c\delta/\sqrt{n}\right), \qquad (۷)$$

برای $n = ۱۰۰, ۵۰۰, ۱۰۰۰$ و $c = ۱/۵$ مقدار λ_n به ترتیب برابر با $F^{-۱}_{O(\mathbf{X},G)}(۰/۹۸۵)$، $F^{-۱}_{O(\mathbf{X},G)}(۰/۹۹۳)$ و $F^{-۱}_{O(\mathbf{X},G)}(۰/۹۹۵)$ است. در عمل به دلیل نامعلوم بودن توزیع F و تابع توزیع تابع دورافتادگی، حتی پس از تعیین مقادیر δ و c نمی‌توان از رابطه (۷) برای تعیین آستانه λ_n استفاده کرد. بدین منظور $\mathrm{out}(\lambda_n,F)$ با شناساگر دورافتادگی بر پایه $\mathbf{X}_۱,\ldots,\mathbf{X}_n$ (یا ناحیهٔ دورافتادگی نمونه) به صورت

[2] True positive rate

$\mathbf{x} \in \mathcal{R}^p$ $OR(\lambda_n, F_n) = \{\mathbf{x} : O(\mathbf{x}, F_n) > \lambda_n\}$، برآورد می‌شود. در این‌صورت دورافتادگی نسبت به نمونه و مقدار آستانهٔ تعیین شدهٔ λ_n که بر اساس نمونه حاصل می‌شود، تعیین می‌شود.

الگوریتم ۱. الگوریتم تعیین آستانه λ_n :
فرض کنید $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ نمونه تحت بررسی باشد و $\hat{G}_n$ نشان دهندهٔ تابع توزیع تجربی آن باشد.
گام ۱- مقادیر دورافتادگی نمونه‌ای $O(\boldsymbol{X}_1, \hat{G}_n), \ldots, O(\boldsymbol{X}_n, \hat{G}_n)$ محاسبه شود.
گام ۲- بر اساس δ و c که از قبل تعیین شده اند، مقدار $1 - \delta\frac{c}{\sqrt{n}}$ محاسبه شود.
گام ۳- بر اساس تابع توزیع تجربی $O(\boldsymbol{X}_1, \hat{G}_n), \ldots, O(\boldsymbol{X}_n, \hat{G}_n)$ چندک $1 - \delta\frac{c}{\sqrt{n}}$ محاسبه شود که این مقدار برابر با λ_n است.

روش شناسائی نقاط دورافتاده که بیان شد، ناوردای آفین است. به عبارت دیگر، بر اساس این روش، نقطه دورافتاده تحت هرگونه تبدیل از محورهای مختصات کماکان به‌عنوان دورافتاده شناسایی می‌شود.

گزاره ۱. اگر تابع ژرفای به کار رفته در تعریف تابع دورافتادگی، در ویژگی ناوردایی آفین صدق کند، آنگاه نقطه‌ای که تحت الگوریتم بالا دورافتاده شناسایی شود، تحت هرگونه تبدیل از محورهای مختصات کماکان به‌عنوان دورافتاده شناسایی می‌شود.

برهان: فرض کنید $\mathbf{X}_1, \ldots, \mathbf{X}_n$ نمونه تحت بررسی باشد، و نقطه $\mathbf{X}_j$ تحت الگوریتم ۱ به عنوان دورافتاده شناسائی شود. برای $i = 1, \ldots, n$ تبدیل $\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \mathbf{b}$، که در آن $\mathbf{A}$ ماتریس $p \times p$ نامنفرد و $\mathbf{b}$ بردار p بعدی دلخواه است، در نظر گرفته می‌شود. چنانچه نشان دهیم $\mathbf{Y}_j$ نیز بر اساس نمونه $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ و تحت الگوریتم ۱ به عنوان دورافتاده شناسائی می‌شود، گزاره اثبات شده است. از ویژگی ناوردایی آفین تابع ژرفا، رابطه $O(\mathbf{X}_i, \hat{G}_n) = O(\mathbf{Y}_i, \hat{U}_n)$، به ازای $i = 1, \ldots, n$ نتیجه خواهد شد، که در آن $\hat{U}_n$ تابع توزیع تجربی $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ است. بر اساس الگوریتم ۱ مقدار آستانه بر اساس $\mathbf{Y}_i$ها نیز برابر با λ_n است. در نتیجه $\mathbf{Y}_j$ به عنوان دورافتاده شناسائی می‌شود.

۴ مطالعه شبیه‌سازی

در این بخش عملکرد روش بیان شده در بخش قبل بر اساس توابع ژرفای مختلف مورد ارزیابی و مقایسه قرار می‌گیرد. شناسایی نقاط دورافتاده بر اساس توابع ژرفای ماهالانوبیس بر اساس برآوردگر بردار میانگین

نمونه‌ای و ماتریس واریانس کوواریانس نمونه‌ای، ژرفای ماهالانوبیس بر اساس برآوردگرهای استوار میانه و میانه قدرمطلق انحراف نقاط از میانه، فضائی، تصویرمبنا، نیم‌فضا و سادکی محاسبه شده و به ترتیب با MO، RMO، SO، PO، HO و LO نشان داده می‌شوند. عملکرد این شناساگرها با دو شناساگر رقیب معرفی شده توسط راک و وودراف (۱۹۹۶) و سرفلینگ و موزوندر (۲۰۱۳) که به ترتیب با MD و $RMSP$ نشان داده می‌شوند، مقایسه می‌شود. همان‌گونه که در بخش مقدمه بیان شد، MD بر اساس فاصله ماهالانوبیس تعریف شده و منحنی تراز بیضوی برای مشاهدات در نظر می‌گیرد. از طرف دیگر، $RMSP$ با تعمیم MD و به کارگیری روش جستجوی تصویر تعریف شده است.

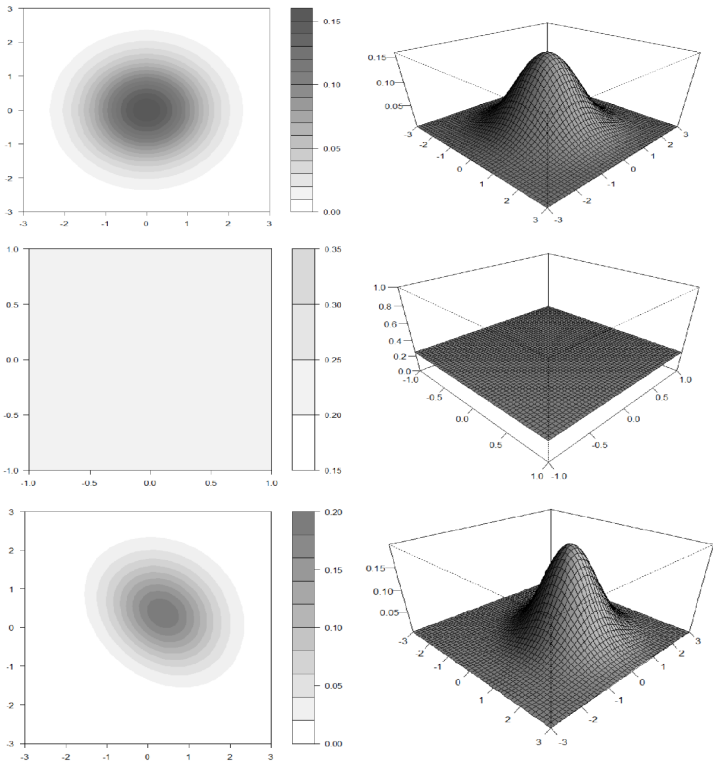
از توزیع نرمال استاندارد دومتغیره به عنوان یک توزیع متقارن بیضوی، توزیع یکنواخت دو متغیره بر مربع $[۰,۱]^۲$ به عنوان یک توزیع متقارن مرکزی که تقارن بیضوی ندارد، و توزیع چوله نرمال دومتغیره استاندارد با پارامتر چولگی ۰٫۲۵ در هر دو مؤلفه، نمونه‌هایی با اندازه ۱۰۰ تولید می‌شود. به منظور درک شهودی بهتر این توزیع‌ها، نمودار منحنی تراز و تابع چگالی آن‌ها در شکل ۱ ارائه شده است. یک مدل آلودگی با $\delta = ۰٫۱$ و $c = ۱٫۵$ در نظر گرفته می‌شود. به عبارت دیگر با انتخاب این مقادیر، نرخ مثبت درست تقریبی تحت مدل آلوده و نرخ مثبت غلط تقریبی تحت مدل غیرآلوده به ترتیب برابر با $\varepsilon_{۱۰۰} = ۰٫۱۵$ و $\alpha_{۱۰۰} = ۰٫۰۱۵$ حاصل می‌شوند. بنابراین مقدار آستانه تجربی به صورت چندک ۰٫۹۸۵ توزیع تجربی تابع دورافتادگی نمونه‌های اولیه تولید شده است. دو الگو برای آلوده کردن ۱۵ درصد از مشاهدات در نظر گرفته می‌شود.

الگوی A: مشاهدات $\mathbf{X}_۱, \ldots, \mathbf{X}_n$ با $K\mathbf{X}_۱, \ldots, K\mathbf{X}_n$، که K یک عدد ثابت است، جایگزین می‌شوند. در مطالعات شبیه سازی این بخش مقدار K برابر با ۵ در نظر گرفته می‌شود. این نوع آلوده کردن بر اساس تغییر در مقیاس مشاهدات اعمال می‌شود.

الگوی B: مشاهدات $\mathbf{X}_۱, \ldots, \mathbf{X}_n$ با $\mathbf{X}_۱ + \mathbf{K}, \ldots, \mathbf{X}_n + \mathbf{K}$، که $\mathbf{K}$ یک بردار دوبعدی ثابت است، جایگزین می‌شوند. در مطالعات شبیه‌سازی این بخش K برداری با مؤلفه‌های برابر با ۴ در نظر گرفته می‌شود. این نوع آلوده کردن بر اساس تغییر در مکان مشاهدات اعمال می‌شود.

دو معیار حساسیت[1] و دقت[2] که توسط فن (۲۰۱۶) برای ارزیابی عملکرد روش‌های شناسائی نقاط دورافتاده به‌کار رفته‌اند، برای شناساگرهای به کار رفته در این مقاله محاسبه می‌شوند. بر اساس جدول ۱، حساسیت و دقت به ترتیب برابر با $\frac{a}{n_۱}$ و $\frac{d}{n_۲}$ هستند. به عبارت دیگر، حساسیت و دقت به ترتیب نشان‌دهندهٔ توانائی شناسایی صحیح یک نقطه دورافتاده به عنوان دورافتاده و توانائی شناسائی صحیح

[1] Sensitivity
[2] Specificity



شکل ۱. چپ- نمودار تراز و راست- از بالا به پایین نمودارهای توابع چگالی توزیع‌های نرمال دومتغیره استاندارد، یکنواخت دو متغیره بر مربع $[۰,۱]^۲$ و چوله نرمال دومتغیره استاندارد با پارامتر چولگی ۰٫۲۵.

جدول ۱. محاسبه معیارهای حساسیت و دقت

وضعیت	دورافتاده واقعی	غیردورافتاده واقعی
شناسائی شده به عنوان دورافتاده	a	b
شناسائی شده به عنوان غیردورافتاده	c	d
	$n_۱ = a + c$	$n_۲ = b + d$

یک نقطه غیردورافتاده به عنوان غیردورافتاده، هستند. حال بر اساس موارد بیان شده روند شبیه‌سازی به صورت زیر انجام می‌شود:

۱- از توزیع مورد نظر نمونه‌ای به اندازه ۱۰۰ استخراج می‌شود.

۲- اگر $\hat{G}_n$ نشان دهندهٔ تابع توزیع تجربی نمونه باشد. آنگاه $O(\mathbf{X}_۱, \hat{G}_n), \ldots, O(\mathbf{X}_n, \hat{G}_n)$ بر اساس تابع دورافتادگی موردنظر محاسبه می‌شوند.

۳- چندک ۰٫۹۸۵ نمونه‌ای $O(\mathbf{X}_1, \hat{G}_n), \ldots, O(\mathbf{X}_n, \hat{G}_n)$ محاسبه می‌شود که این مقدار برابر با λ_n است.
۴- الگوی آلودگی مورد نظر (الگوی A یا B) برای جایگذاری ۱۵ مشاهده آخر از مجموعه نمونه ۱۰۰ تایی تولید شده در نظر گرفته می‌شود.
۵- تابع دورافتادگی برای هر عضو نمونه (کل اعضا شامل آلوده شده و غیرآلوده) محاسبه شده و با مقایسه مقدار دورافتادگی هر عضو با آستانه موردنظر، معیارهای حساسیت و دقت حاصل می‌شوند.
۶- مراحل ۱ تا ۵، ۱۰۰۰ بار تکرار می‌شود و میانگین موارد حاصل شده در مرحله ۴ به عنوان برآورد نهایی معیارهای حساسیت و دقت گزارش می‌شوند.

نتایج در جدول ۲ برای توزیع نرمال استاندارد دومتغیره، توزیع یکنواخت دو متغیره بر مربع $[۰, ۱]^۲$ و توزیع چوله نرمال دومتغیره استاندارد با پارامتر چولگی ۰٫۲۵ در هر دو مؤلفه، آمده است. $\mathbb{X}_n$، $\mathbb{X}_n^A$ و $\mathbb{X}_n^B$ به ترتیب نمونه اصلی، نمونه آلوده شده تحت الگوی A و نمونه آلوده شده تحت الگوی B هستند. برای ستون‌های $\mathbb{X}_n$ انتظار می‌رود، مقدار حساسیت حدود ۰٫۰۱۵ و مقدار دقت حدود ۰٫۹۸۵ باشد، که بررسی نتایج، عملکرد قابل قبول همه شناساگرها بجز شناساگرهای مبتنی توابع دورافتادگی نیم‌فضا و سادکی را نشان می‌دهد. برای $\mathbb{X}_n^A$ و $\mathbb{X}_n^B$ واضح است، هرچه مقادیر حساسیت و دقت به یک نزدیک‌تر باشند، نشان دهندهٔ عملکرد بهتر روش موردنظر است. همانطورکه ملاحظه می‌شود، عملکرد شناساگرهای مبتنی بر توابع دورافتادگی، تحت تاثیر تابع ژرفای به کار رفته در آن‌ها است. واضح است تغییر و آلودگی در داده‌ها، بیشترین تاثیر را بر روی روش‌هایی که مبتنی بر برآوردگرهای استوار نیستند، مانند MO، SO و MD که بر اساس برآوردگرهای نااستوار پارامتر مکان و پراکندگی تعریف می‌شوند، دارد. نتایج نشان می‌دهند، همان‌گونه که انتظار می‌رود، شناساگر رقیب MD که بر اساس فاصله ماهالانوبیس تعریف می‌شود، عملکرد مشابه با MO دارد. از طرف دیگر، برای توزیع‌های غیربیضوی یکنواخت و چوله نرمال، $RMSP$ عملکرد بهتری نسبت به MD دارد. از طرف دیگر ژرفای سادکی و نیم‌فضا بر اساس موقعیت نسبی یک نقطه نسبت به تابع توزیع تعریف می‌شوند. بنابراین شناساگرهای تعریف شده بر اساس آن‌ها ساختار هندسی نمونه را در نظر می‌گیرند، اما عملکرد مناسبی، در شناسائی صحیح دورافتادگی و یا عدم دورافتادگی نقاط ندارند.

شناساگرهای معرفی شده بر اساس تابع ژرفای تصویرمبنا، PO، و تابع ژرفای ماهالانوبیس با برآوردگرهای استوار،RMO ، تحت هر دو الگوی آلودگی A و B عملکرد بسیار بهتری نسبت به سایر شناساگرها دارند. همچنین لازم به ذکر است، برخلاف توابع ژرفای سادکی و نیم‌فضا، محاسبه توابع ژرفای

جدول ۲. حساسیت و دقت تجربی برای توزیع‌های دو متعیره مختلف و مدل‌های آلودگی A و B

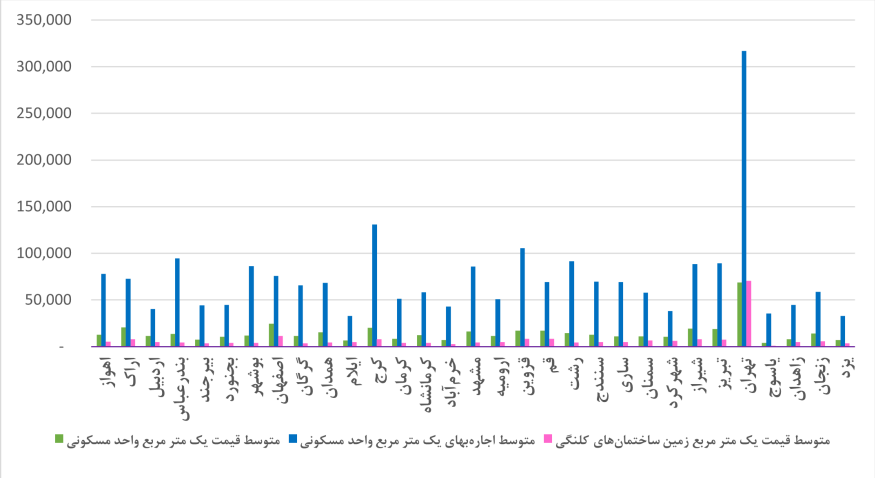
توزیع دو متغیره	تابع دورافتادگی	حساسیت			دقت		
		$\mathbb{X}_n$	$\mathbb{X}_n^A$	$\mathbb{X}_n^B$	$\mathbb{X}_n$	$\mathbb{X}_n^A$	$\mathbb{X}_n^B$
نرمال	MO	۰٫۰۱۷	۰٫۵۰۴	۰٫۹۱۰	۰٫۹۸۰	۱	۰٫۹۹۴
	RMO	۰٫۰۱۷	۰٫۷۸۹	۰٫۹۱۸	۰٫۹۸۰	۰٫۸۹۶	۰٫۹۸۸
	SO	۰٫۰۱۷	۰٫۵۳۳	۰٫۷۳۰	۰٫۹۸۰	۱	۰٫۹۸۴
	PO	۰٫۰۱۸	۰٫۷۹۶	۰٫۹۵۴	۰٫۹۸۰	۰٫۸۹۶	۰٫۹۹۲
	HO	۰٫۰۸۷	۰٫۴۱۰	۰٫۲۴۴	۰٫۹۱۰	۰٫۸۹۹	۰٫۹۳۶
	LO	۰٫۰۸۷	۰٫۴۱۰	۰٫۲۴۴	۰٫۹۱۰	۰٫۸۹۹	۰٫۹۳۵
	MD	۰٫۰۱۷	۰٫۵۰۹	۰٫۹۱۱	۰٫۹۸۰	۱	۰٫۹۹۴
	$RMSP$	۰٫۰۱۶	۰٫۵۰۲	۰٫۹۰۹	۰٫۹۸۰	۱	۰٫۹۹۴
یکنواخت	MO	۰٫۰۲۲	۰٫۷۶۶	۰٫۷۴۹	۰٫۹۸۰	۱	۰٫۹۸۷
	RMO	۰٫۰۲۲	۰٫۹۲۰	۱	۰٫۹۸۰	۰٫۸۹۹	۰٫۸۳۵
	SO	۰٫۰۲۲	۰٫۸۰۹	۰٫۲۵۰	۰٫۹۸۰	۱	۰٫۹۵۰
	PO	۰٫۰۲۱	۰٫۹۱۶	۱	۰٫۹۸۰	۰٫۸۹۹	۰٫۹۸۴
	HO	۰٫۱۱۸	۰٫۴۶۳	۰٫۲۹۲	۰٫۸۸۲	۱	۰٫۹۲۰
	LO	۰٫۱۱۸	۰٫۴۶۴	۰٫۲۹۳	۰٫۸۸۱	۱	۰٫۹۲۰
	MD	۰٫۰۲۱	۰٫۷۵۶	۰٫۷۳۹	۰٫۹۷۷	۱	۰٫۹۸۱
	$RMSP$	۰٫۰۱۹	۰٫۸۴۶	۰٫۹۵۲	۰٫۹۸۰	۰٫۸۹۹	۰٫۹۸۳
چوله‌نرمال	MO	۰٫۰۲۰	۰٫۵۱۱	۰٫۲۰۰	۰٫۹۸۰	۱	۰٫۹۹۴
	RMO	۰٫۰۲۰	۰٫۷۹۶	۰٫۹۹۵	۰٫۹۸۰	۰٫۸۹۶	۰٫۹۸۶
	SO	۰٫۰۱۹	۰٫۵۳۶	۰٫۰۷۳	۰٫۹۸۰	۱	۰٫۹۸۴
	PO	۰٫۰۲۱	۰٫۸۰۳	۰٫۹۷۹	۰٫۹۸۰	۰٫۸۹۶	۰٫۹۸۴
	HO	۰٫۰۸۸	۰٫۴۰۷	۰٫۲۴۶	۰٫۹۱۱	۰٫۸۹۸	۰٫۹۳۶
	LO	۰٫۰۸۸	۰٫۴۰۶	۰٫۲۴۶	۰٫۹۱۱	۰٫۸۹۸	۰٫۹۳۶
	MD	۰٫۰۲۱	۰٫۶۱۱	۰٫۷۹۳	۰٫۹۸۰	۱	۰٫۹۹۰
	$RMSP$	۰٫۰۲۰	۰٫۷۶۳	۰٫۹۲۱	۰٫۸۷۹	۰٫۸۹۹	۰٫۹۸۳

ماهالانوبیس و تصویرمبنا دارای پیچیدگی محاسباتی نیست. بنابراین به طور کلی می‌توان روش ناوردای آفین پیشنهادی بر اساس توابع ژرفای تصویر–مبنا و ماهالانوبیس با برآوردگرهای استوار را از جهات توانائی شناسائی نقاط دورافتاده و مرتبه محاسباتی به عنوان مطلوب‌ترین شناساگرها در مقایسه با سایر شناساگرهای مبتنی بر تابع ژرفا و همچنین شناساگرهای رقیب در نظر گرفت.

۵ تحلیل داده‌های مسکن

مرکز آمار ایران، برآورد پارامترهایی نظیر متوسط قیمت یک مترمربع واحد مسکونی، متوسط اجاره‌بهای یک متر مربع واحد مسکونی و متوسط قیمت یک مترمربع زمین ساختمان‌های مسکونی کلنگی را از طریق طرح‌های نمونه‌گیری در طی چندین دوره زمانی ارائه داده است. ساختمان مسکونی کلنگی، منظور آن دسته از ساختمان‌های مسکونی است که به منظور تخریب کامل بنا و احداث بنای جدید برروی زمین آن، مورد

معامله قرار گرفته است (این دسته از ساختمان‌ها می‌توانند پس از خرید مورد بازسازی و استفاده قرار گیرند). در این بخش به دنبال شناسائی شهرهایی از کشور ایران هستیم که به طور کلی در زمینه ارزش مالی واحد مسکونی متفاوت از سایر شهرها هستند. سه متغیر متوسط قیمت یک مترمربع واحد مسکونی، متوسط اجاره‌بهای یک متر مربع واحد مسکونی و متوسط قیمت یک مترمربع زمین ساختمان‌های کلنگی در سال ۱۳۹۷ به تفکیک شهرهای منتخب اهواز، اراک، اردبیل، بندرعباس، بیرجند، بجنورد، بوشهر، اصفهان، گرگان، همدان، ایلام، کرج، کرمان، کرمانشاه، خرم‌آباد، مشهد، ارومیه، قزوین، قم، رشت، سنندج، ساری، سمنان، شهرکرد، شیراز، تبریز، تهران، یاسوج، زاهدان، زنجان و یزد، در نظر گرفته می‌شوند. در شکل ۲ نمودار داده‌های موردنظر مشاهده می‌شود. همان‌گونه که مشاهده می‌کنید در هر سه متغیر شهر تهران اختلاف



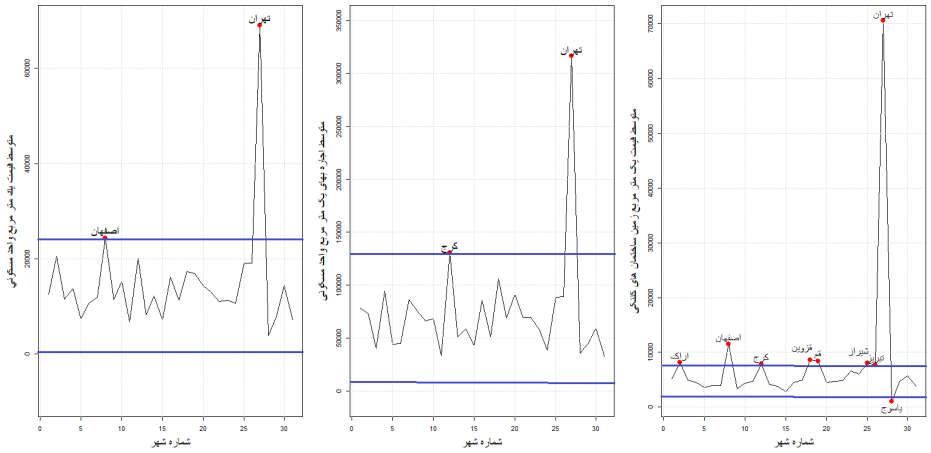
شکل ۲. متوسط قیمت یک مترمربع زمین ساختمان‌های کلنگی، متوسط قیمت اجاره مسکن و متوسط قیمت یک مترمربع واحد مسکونی در سال ۱۳۹۷ به تفکیک شهرهای منتخب

معنی‌داری با سایر شهرها دارد، به‌طوری در مقایسه با این اختلاف، تفاوت میان سایر شهرها ناچیز به نظر می‌رسد. ابتدا، بر اساس روش‌ها و آزمون‌های معروف در حالت یک‌متغیره به شناسائی نقاط دورافتاده می‌پردازیم. ابتدا بر اساس آزمون کولموگروف-اسمیرنوف و شاپیرو-ویلک نرمال بودن متغیرها بررسی شده است که نرمال بودن هیچ‌یک از سه متغیر مورد نظر معنی‌دار نشد. بنابراین با استفاده از روش‌های استوار یک‌متغیره به دنبال شناسائی نقاط دورافتاده یک‌متغیره هستیم. یکی از روش‌های استوار، بازه‌ای با مرکز میانه مشاهدات و کران‌هایی بر اساس میانه انحراف نقاط از میانه است. به عبارت دیگر اگر Med نشان دهنده میانه مشاهدات یک‌متغیره و MAD میانه قدرمطلق انحراف نقاط از میانه باشد، که در بخش ۲ به

ترتیب در معادلات (۲) و (۳) تعریف شده‌اند، آنگاه نقاطی که در خارج از بازه

$$(Med - ۳MAD, Med + ۳MAD) \qquad (۸)$$

قرار می‌گیرند، به عنوان دورافتاده در نظر گرفته می‌شوند. بر اساس ترتیبی که در بالا اسم شهرها آمده است، شماره ۱ تا ۳۱ به آن‌ها اختصاص داده شده است. در شکل ۳ برای هر یک از متغیرها کران پائین و بالای این بازه با خطوط افقی نمایش داده شده است. همانگونه که مشاهده می‌شود، سه شهر تهران، اصفهان و کرج برای حداقل دو متغیر در خارج از بازه قرار گرفته‌اند و به عنوان دورافتاده در نظر گرفته می‌شوند. از طرف دیگر شهر یاسوج برای هر سه متغیر دارای کمترین مقدار در بین شهرها است و برای متغیر متوسط قیمت یک مترمربع زمین ساختمان‌های کلنگی به عنوان نقطه دورافتاده شناسائی شده است. حال آزمون‌های



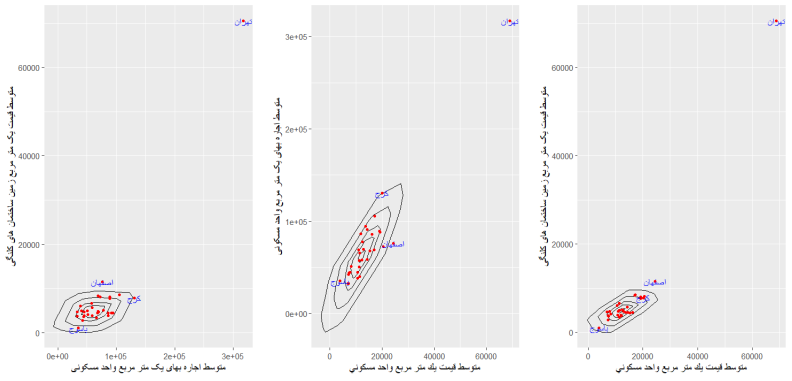
شکل ۳. نمودار پراکندگی به تفکیک متغیرها، خطوط افقی کران های پائین و بالای بازه تعریف شده در (۸) را برای هر متغیر نشان می‌دهند.

معروف ناپارامتری یک‌متغیره شناسائی نقاط دورافتاده، گرابس (۱۹۵۰) و کای‌دو (دیکسون ، ۱۹۵۰)، نیز برای هر یک از متغیرها به کار می‌رود. همانگونه که در جدول ۳ ملاحظه می‌شود، شهر تهران بر اساس هر دو آزمون و برای هر سه متغیر به عنوان دورافتاده شناسائی شده است. بر اساس نتایج می‌توان استدلال کرد که شهرهای اصفهان و یاسوج نیز از نظر دورافتادگی نسبت به سایر شهرها در رتبه‌های بعدی قرار گرفته‌اند.

جدول ۳. شهرهای شناسائی شده به عنوان دورافتاده بر اساس آزمون گرابس و کای‌دو

متغیر	آزمون گرابس	آزمون کای‌دو
متوسط قیمت یک مترمربع واحد مسکونی	تهران	تهران، اصفهان، یاسوج
متوسط اجاره‌بهای یک متر مربع واحد مسکونی	تهران	تهران، کرج، قزوین
متوسط قیمت یک مترمربع زمین ساختمان‌های کلنگی	تهران، اصفهان	تهران، اصفهان، یاسوج

حال روش ناپارامتری چندمتغیره بیان شده در این مقاله بر روی داده‌ها به‌کار برده می‌شود. با توجه به نتایج شبیه‌سازی و عملکرد مناسب تابع ژرفای تصویر–مبنا، تابع دورافتادگی بر اساس آن محاسبه می‌شود. در شکل ۴، خطوط تراز که در معادله (۴) تعریف شده‌اند، بر اساس مقدار تابع دورافتادگی تصویر–مبنا برای داده‌های دومتغیره و مقادیر t برابر با ۰٫۷۷، ۰٫۷، ۰٫۶، ۰٫۵ و ۰٫۳ برای هر یک از دو متغیر ممکن رسم شده است. همان‌گونه که مشاهده می‌شود، شهرهای تهران و اصفهان در خارج از منحنی تراز ۰٫۷۷ و شهر کرج نیز بر روی این منحنی قرار دارد. همچنین مقدار تابع دورافتادگی تصویر–مبنا برای داده‌های سه‌متغیره محاسبه شده است و مقادیر آن در جدول ۴ آمده است. با در نظر گرفتن نرخ مثبت نادرست ۰٫۰۵، مقدار آستانه تجربی به صورت چندک ۰٫۹۵ توزیع تجربی مقادیر دورافتادگی به دست آمده است. چندک تجربی ۰٫۹۵ مقادیر دورافتادگی برابر با ۰٫۸۶۱ است. بنابراین بر اساس مقادیر دورافتادگی شهرهای تهران و اصفهان از نظر ارزش مالی مسکن نسبت به سایر شهرهای کشور، دورافتاده محسوب می‌شوند.



شکل ۴. منحنی تراز ۰٫۷۷، ۰٫۷، ۰٫۶، ۰٫۵ و ۰٫۳ بر اساس تابع دورافتادگی تصویر–مبنا

جدول ۴. مقدار تابع دورافتادگی تصویر–مبنا

شهر	تهران	اصفهان	کرج	اراک	قزوین	یاسوج	قم	بندعباس	شیراز	بوشهر	تبریز
PO	۰٫۹۸۴	۰٫۸۷۹	۰٫۸۴۶	۰٫۸۰۱	۰٫۸۱۰	۰٫۷۹۵	۰٫۷۸۷	۰٫۷۶۸	۰٫۷۵۷	۰٫۷۵۴	۰٫۷۴۶
شهر	رشت	همدان	مشهد	ایلام	خرم آباد	شهرکرد	اردبیل	یزد	زاهدان	سمنان	بیرجند
PO	۰٫۷۴۵	۰٫۷۴۲	۰٫۷۳۴	۰٫۷۰۹	۰٫۶۹۲	۰٫۶۸۷	۰٫۶۸۵	۰٫۶۶۸	۰٫۶۶۷	۰٫۶۶۳	۰٫۶۵۷
شهر	اهواز	زنجان	کرمانشاه	بجنورد	گرگان	کرمان	ساری	سنندج	ارومیه		
PO	۰٫۶۴۷	۰٫۶۴۳	۰٫۶۳۹	۰٫۶۲۹	۰٫۶۱۷	۰٫۶۰۳	۰٫۵۹۴	۰٫۵۰۰	۰٫۴۷۹		

۶ بحث و نتیجه‌گیری

در این مقاله، یک روش برای شناسائی نقاط دورافتاده چندمتغیره بر اساس تابع ژرفا معرفی شد. روش پیشنهادی دارای ویژگی مطلوب ناوردای آفین است و پیاده‌سازی آن نسبت به بیشتر روش‌های چندمتغیره که دارای پیچیدگی محاسباتی هستند، ساده‌تر است. بر اساس مطالعات شبیه‌سازی، عملکرد روش پیشنهادی بر اساس توابع ژرفای مختلف مورد بررسی قرار گرفت و همچنین با شناساگرهای رقیب مقایسه شد. نتایج شبیه‌سازی نشان داد، روش پیشنهادی بر اساس توابع ژرفای تصویر–مبنا و ماهالانوبیس با برآوردگرهای استوار، دارای عملکرد مطلوب‌تری در مقایسه با سایر شناساگرها است. در ادامه، روش پیشنهادی بر اساس تابع ژرفای تصویر–مبنا برای داده‌های ارزش مالی واحد مسکونی ایران در سال ۱۳۹۷ به کار رفت. نتایج نشان داد، شهرهای تهران و اصفهان را می‌توان از نظر ارزش مالی واحد مسکونی متفاوت از سایر شهرهای ایران در نظر گرفت.

مراجع

روزبه، م. و امینی، م. (۱۳۹۸)، برآوردگر استوار مرزبندی شده تعمیم‌یافته محتمل در مدل رگرسیون نیمه‌پارامتری، مجله علوم آماری، (۲)۱۳، ۴۶۰-۴۴۱.

روزبه، م. و معنوی، م. (۱۳۹۹)، مدل‌سازی سن تقویمی به روش رگرسیون ستیغی کمترین توان‌های دوم پیراسته، مجله علوم آماری، (۲)۱۴، ۴۲۸-۴۰۹.

Barnett, V. and Lewis, T. (1994), *Outliers in Statistical Data*, John Wiley & Sons.

Chen, Y., Dang, X., Peng, H. and Bart, L. (2008), Outlier Detection with

the Kernelized Spatial Depth Function, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(2), 288-305.

Dang, X. and Serfling, R. (2010), Nonparametric Depth-based Multivariate Outlier Identifiers, and Masking Robustness Properties, *Journal of Statistical Planning and Inference*, **140**(1), 198–213.

Dehghan, S. and Faridrohani, M. R. (2019), Affine Invariant Depth-Based Tests for the Multivariate One-Sample Location Problem, *Test*, **28**(3), 671-693.

Dixon, W. J. (1950), Analysis of Extreme Values. *The Annals of Mathematical Statistics*, **21**(4), 488-506.

Dovoedo, Y. H. and Chakraborti, S. (2013), Outlier Detection for Multivariate Skew-Normal Data: a Comparative Study, *Journal of Statistical Computation and Simulation*, **83**(4), 773-783.

Fan, Yi. (2016), *New Nonparametric Approaches for Multivariate and Functional Data Analysis in Outlier Detection, Construction of Tolerance Tubes, and Clustering*, Diss. Rutgers University-Graduate School-New Brunswick.

Grubbs, F. E. (1950), Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, **21**(1), 27-58.

Liu, R. Y. (1988), On a Notion of Simplicial Depth, *Proceedings of the National Academy of Sciences*, **85**(6), 1732–1734.

Liu, R. Y. and Singh, K. (1993), A Quality Index Based on Data Depth and Multivariate Rank Tests, *Journal of the American Statistical Association*, **88**(421), 252–260.

Hawkins, D. M. (1980), *Identification of Outliers*, Chapman and Hall.

Kannan, K., Senthamarai, and Manoj, K. (2015), Outlier Detection in Multivariate Data, *Applied Mathematical Sciences*, **9**(47), 2317-2324.

Miller, J. (1991), Reaction Time Analysis with Outlier Exclusion: Bias Varies with Sample Size. *The Quarterly Journal of Experimental Psychology*, **43**(4), 907–912.

Serfling, R. and Mazumder, S. (2013), Computationally Easy Outlier Detection via Projection Pursuit with Finitely Many Directions, *Journal of Nonparametric Statistics*, **25**(2), 447-461.

Rocke, D. M. and Woodruff, D. L. (1996), Identification of Outliers in Multivariate Data, *Journal of the American Statistical Association*, **91**(435), 1047-1061.

Tukey, J. W. (1975), Mathematics and the Picturing of Data, *Proceedings of the international congress of mathematicians*, **2**, 523–531.

Vardi, Y. and Zhang, C. H. (2000), The Multivariate L1-Median and Associated Data Depth, *Proceedings of the National Academy of Sciences*, **97**(4), 1423-1426.

Zuo, Y. and Serfling, R. (2000), General Notions of Statistical Depth Function, *Annals of statistics*, **28**(2), 461–482.

Zuo, Y. (2003), Projection-based Depth Functions and Associated Medians. *The Annals of Statistics*, **31**(5) 1460–1490.

Journal of Statistical Sciences, Autumn and Winter, 2021
Vol. 15, No. 2, pp 443-462
DOI: 10.29252/jss.15.2.443

Multivariate Outlier Detection Based on Depth–Based Outlyingness Function

Dehghan, S., Faridrohani, M. R.
Department of Statistics, Faculty of Mathematical Sciences, University of Shahid Beheshti, Tehran, Iran.

Abstract: The concept of data depth has provided a helpful tool for nonparametric multivariate statistical inference by taking into account the geometry of the multivariate data and ordering them. Indeed, depth functions provide a natural centre-outward order of multivariate points relative to a multivariate distribution or a given sample. Since the outlingness of issues is inevitably related to data ranks, the centre-outward ordering could provide an algorithm for outlier detection. In this paper, based on the data depth concept, an affine invariant method is defined to identify outlier observations. The affine invariance property ensures that the identification of outlier points does not depend on the underlying coordinate system and measurement scales. This method is easier to implement than most other multivariate methods. Based on the simulation studies, the performance of the proposed method based on different depth functions has been studied. Finally, the described method is applied to the residential houses' financial values of some cities of Iran in ۲۰۱۸.

Keywords: Depth function, Outlyingness function, Outlier, Affine invariance.

Mathematics Subject Classification (2010): 62G99, 62H99.