آزمون همزمان استقلال برای زیربردارهای چند بردار با بُعد نسبتاً بالای نرمال چندمتغیره

داریوش نجارزاده

گروه آمار، دانشکده علوم ریاضی، دانشگاه تبریز

چکیده: آزمون فرض استقلال میان زیربردارهای یک بردار p-متغیره، به عنوان پیش‌نیاز بسیاری از آزمون‌های آماری، همواره مورد توجه بوده است. وقتی اندازه نمونه n در مقایسه با بُعد p خیلی بزرگ است، آزمون نسبت درستنمایی با توزیع تقریبی خی‌دو، عملکرد قابل قبولی دارد. برای "داده‌های با بُعد نسبتاً بالا" که در آنها n در قیاس با p چندان بزرگ نیست، تقریب خی‌دو برای توزیع آماره آزمون نسبت درستنمایی کارایی لازم را ندارد. به عنوان یک حالت جامع‌تر، در این مقاله، آزمونی همزمان در k جامعه p-متغیره نرمال با بُعد نسبتاً بالا که در هر جامعه آزمون استقلال میان زیربردارهای دلخواه آزموده می‌شود، مد نظر قرار گرفته است. به منظور آزمون این فرض، یک تقریب نرمال برای توزیع آماره آزمون نسبت درستنمایی تحت فرض صفر بدست آمده است. علاوه بر این، به منظور تصدیق عملکرد بهتر تقریب نرمال پیشنهادی بر تقریب خی‌دوی کلاسیک، مطالعه شبیه‌سازی انجام شده است. در پایان، کاربردی از روش پیشنهادی بر مجموعه داده سرطان پرستات ارائه شده است.

واژه‌های کلیدی: توزیع نرمال چندمتغیره، آزمون نسبت درست‌نمایی، داده‌های با بُعد نسبتاً بالا، آزمون استقلال، تابع گامای چندمتغیره.

---

* آدرس الکترونیک مسئول مقاله: داریوش نجارزاده، d_najarzadeh@tabrizu.ac.ir

*کد موضوع‌بندی ریاضی (۲۰۱۰): 62H15، 62H10، 60F05.

۱ مقدمه

در تحلیل‌های چند متغیره در حالت یک جامعه ($k = ۱$)، آزمون استقلال زیربردارهای یک بردار p-متغیره نرمال با بعد بالا همواره پیش‌نیاز بسیاری از آزمون‌ها است. در بسیاری از مواقع لازم است همین آزمون به طور همزمان روی چندین ($k > ۱$) جامعه p-متغیره نرمال اجرا شود. کاربردهایی از این نوع آزمون استقلال چه در حالت تک جامعه و چه در حالت چندین جامعه بر داده‌های ریزآرایه[۱]، داده‌های مالی[۲]، داده‌های مصرف[۳]، داده‌های ساخت پیشرفته[۴] و داده‌های چندرسانه‌ای[۵] و غیره مشهود است (مائو، ۲۰۱۸؛ چن و همکاران، ۲۰۱۸؛ لئونگ و درتون، ۲۰۱۸). به عنوان مثال، در تحلیل داده‌های ریزآرایه بررسی اینکه بین ژن‌های مختلف استقلال وجود دارد یا نه، همواره از اهمیت خاصی برای تحلیل‌های آتی برخوردار است.

برای آزمون استقلال زیربردارها در حالت $k > ۱$، روش آزمون نسبت درست‌نمایی[۶] (LRT) با توزیع مجانبی خی‌دو پیشنهاد شده است. ویلکس (۱۹۳۸) نشان داد برای اندازه‌های نمونه‌ای $n_۱, \ldots, n_k$ بزرگتر از بُعد p با نسبت‌های $\frac{p}{n_i}$ نزدیک به صفر، توزیع تحت فرض صفر آماره LRT، به توزیع مجانبی خی‌دو ($\chi^۲$) همگراست. در حالت "داده‌های با بُعد نسبتاً بالا"[۷]، با این تعریف که در آن n_iها از بُعد p بزرگ‌اند و نسبت‌های $\frac{p}{n_i}$ اعدادی نزدیک به یک هستند و همچنین داده‌های با بعد بالا، با این تعریف که در آن n_iها از بُعد p کوچک‌اند، تقریب خی‌دو به آزمونی با اندازه[۸] بسیار بزرگتر از سطح معنی‌داری اسمی α یا به طور معادل آزمونی با اندازه‌ای متورم منجر می‌شود. این ایراد در مورد تقریب خی‌دوی آماره LRT در آزمون فرض‌های دیگر نیز مشاهده شده است، که از این جمله می‌توان به آزمون فرض برابری ماتریس کوواریانس با ماتریس همانی یا همان آزمون کُرویت[۹] برای داده‌های با بُعد نسبتاً بالا درکار بای و همکاران (۲۰۰۹) اشاره کرد که در آن برای رفع مشکل متورم بودن اندازه آزمون

[1] Microarray data
[2] Financial data
[3] Consumer data
[4] Modern manufacturing data
[5] Multimedia data
[6] Likelihood Ratio Test
[7] Moderately high dimensional data
[8] Test size
[9] Sphericity test

LRT به کمک نظریه ماتریس‌های تصادفی و آماره‌های طیفی خطی[10] روشی برای تصحیح این آماره ارائه شده است. موارد مشابه دیگری را می‌توان در پژوهش‌های اسکات (۲۰۰۵، ۲۰۰۷)، لدویت و ولف (۲۰۰۲)، چن و همکاران (۲۰۱۰)، ژیانگ و همکاران (۲۰۱۲)، ژیانگ و یانگ (۲۰۱۳) یافت. روش کلاسیک LRT برای آزمون استقلال میان زیربردارهای $k = ۱$ بردار p-متغیره نرمال با بُعد نسبتاً بالا، قبلاً توسط ژیانگ و یانگ (۲۰۱۳) مورد مطالعه قرار گرفته است. آنها نشان دادند که برای حالتی که بُعد p متناسب با اندازه نمونه n به صورت $\frac{p}{n} \longrightarrow y \in (۰, ۱]$ رشد می‌کند، تقریب خی‌دو عملاً قابل استفاده نبوده و در این حالت توزیع تحت فرض صفر آماره LRT به جای توزیع خی‌دو به یک توزیع نرمال همگرا است.

به عنوان توسیعی از یافته‌های ژیانگ و یانگ (۲۰۱۳) به حالت بیش از یک جامعه، در این مقاله، آزمون همزمان استقلال میان زیربردارهای $k > ۱$ بردار p-متغیره نرمال با بُعد نسبتاً بالا مورد مطالعه قرار گرفته است. به بیان دیگر، با این فرض که به ازای هر $i = ۱, \ldots, k$، بردار تصادفی $\boldsymbol{X}_i$ دارای توزیع نرمال p-متغیره با بردار میانگین $\boldsymbol{\mu_i}$ و کوواریانس $\boldsymbol{\Sigma}_i$، یا به اختصار توزیع $N_p(\boldsymbol{\mu_i}, \boldsymbol{\Sigma}_i)$ است و $\boldsymbol{X}_i' = (\boldsymbol{X}_۱^{(i)\prime}, \cdots, \boldsymbol{X}_{k_i}^{(i)\prime})'$ افرازی از $\boldsymbol{X}_i$ به $۱ \leq k_i \leq p$ زیربردار باشد، هدف این مقاله آزمون استقلال زیربردارهای $\boldsymbol{X}_۱^{(i)}$، ...، $\boldsymbol{X}_{k_i}^{(i)}$ به طور همزمان در همه k جامعه بر اساس نمونه‌های تصادفی با اندازه‌های نمونه‌ای $n_i > p + ۱$، $i = ۱, \ldots, k$ با نسبت‌های $\frac{p}{n_i}$ها نزدیک به یک است. در این حالت، ثابت می‌شود که توزیع آماره LRT به یک توزیع نرمال همگرا خواهد شد. توجه شود که حالت داده‌های با بعد بالا در این مقاله مورد مطالعه نیست.

در بخش ۲، ضمن استخراج آماره LRT، همگرایی توزیع این آماره به یک توزیع نرمال، با میانگین و واریانسی خوش‌تعریف، اثبات شده است. در بخش ۳، یک مطالعه شبیه‌سازی به منظور بررسی اندازه و توان آزمون‌های مورد مطالعه انجام شده است. نتایج شبیه‌سازی حاکی از این است که آزمون معرفی شده در این مقاله بر آزمون کلاسیک خی‌دو برتری دارد. مثالی کاربردی از روش پیشنهادی روی مجموعه داده سرطان پرستات در بخش ۴ بررسی می‌شود. در بخش ۵ به بحث و نتیجه‌گیری پرداخته می‌شود.

[10] Linear spectral statistics

۲ بخش نظری

فرض کنید بردار $\boldsymbol{X}_i$ دارای توزیع $N_p(\boldsymbol{\mu_i}, \boldsymbol{\Sigma}_i)$ است. اگر $\boldsymbol{X}_i$ به $1 \leq k_i \leq p$ زیربردار به صورت $\boldsymbol{X}_i' = (\boldsymbol{X}_1^{(i)\prime}, \cdots, \boldsymbol{X}_{k_i}^{(i)\prime})'$ و $\boldsymbol{X}_r^{(i)} \in \mathbb{R}^{p_r^{(i)}}$ افراز شود، آنگاه $(p_1^{(i)}, \ldots, p_{k_i}^{(i)})$ افرازی از p با $p = \sum_{r=1}^{k_i} p_r^{(i)}$ است. متناظر با افراز $\boldsymbol{X}_i$ بردار میانگین $\boldsymbol{\mu_i}$ و ماتریس کوواریانس $\boldsymbol{\Sigma}_i$ به ترتیب به صورت $\boldsymbol{\mu}_i' = (\boldsymbol{\mu}_1^{(i)\prime}, \ldots, \boldsymbol{\mu}_{k_i}^{(i)\prime})'$ و $\boldsymbol{\Sigma}_i := (\boldsymbol{\Sigma}_{l\times m}^{(i)})_{k_i \times k_i}$ افراز خواهند شد، که در آن $\boldsymbol{\Sigma}_{\ell\times m}^{(i)} = \mathrm{Cov}(\boldsymbol{X}_\ell^{(i)}, \boldsymbol{X}_m^{(i)})$ درایه ماتریسی روی سطر ℓام و ستون mام ماتریس بلوکی حاصل از افراز $\boldsymbol{\Sigma}_i$ است. با این قراردادها، فرض استقلال زیربردارهای $\boldsymbol{X}_1^{(i)}, \ldots, \boldsymbol{X}_{k_i}^{(i)}$، $i = 1, \ldots, k$ یا همان $H_\circ$، فرضی است که در این مقاله آزمون می‌شود. این فرض را به شکل معادل می‌توان به صورت قابلیت نوشتن چگالی $\boldsymbol{X}_i$ به شکل حاصل‌ضرب چگالی‌های $\boldsymbol{X}_1^{(i)}, \ldots, \boldsymbol{X}_{k_i}^{(i)}$ نیز بیان کرد. از این رو می‌توان فرض $H_\circ$ را به صورت

$$H_\circ : f_{\boldsymbol{X}_i}(\boldsymbol{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \prod_{r=1}^{k_i} f_{\boldsymbol{X}_r^{(i)}}(\boldsymbol{x}_r^{(i)}; \boldsymbol{\mu}_r^{(i)}, \boldsymbol{\Sigma}_{rr}^{(i)}),\ i = 1, \ldots, k, \qquad (1)$$

نیز نوشت. حال، فرض کنید به ازای هر $i = 1, \cdots, k$، بردارهای $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$ نمونه‌ای تصادفی با اندازه n_i از $\boldsymbol{X}_i$ هستند. بنابراین تحت فرض صفر (۱) تابع درست‌نمایی این مشاهدات به صورت

$$L(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} f_{\boldsymbol{X}_{ij}}(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

است. می‌توان نشان داد (اندرسون، ۲۰۰۳) که

$$\sup_{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, i=1,2,\ldots k} L(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k) = \prod_{i=1}^{k} (2\pi e n_i^{-1})^{-\frac{n_i p}{2}} |\boldsymbol{W}_i|^{-\frac{n_i}{2}},$$

به گونه‌ای که $\boldsymbol{W}_i = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)' := (\boldsymbol{W}_{\ell\times m}^{(i)})_{k_i \times k_i}$. به سادگی می‌توان دید که تحت فرض $H_\circ$ در (۱)، ماتریس کوواریانس $\boldsymbol{\Sigma}_i$ برابر ماتریس قطری بلوکی $\boldsymbol{\Sigma}_i^{H_\circ}$ به صورت

$$\boldsymbol{\Sigma}_i^{H_\circ} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{(i)} & \circ & \cdots & \circ \\ \circ & \boldsymbol{\Sigma}_{22}^{(i)} & \cdots & \circ \\ \vdots & \vdots & \ddots & \circ \\ \circ & \circ & \cdots & \boldsymbol{\Sigma}_{k_i k_i}^{(i)} \end{bmatrix}.$$

خواهد بود. بنابراین، تحت فرض $H_۰$ در (۱)،

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i^{H_۰},i=۱,\dots k} L(\boldsymbol{\mu}_۱,\dots,\boldsymbol{\mu}_k;\boldsymbol{\Sigma}_۱^{H_۰},\dots,\boldsymbol{\Sigma}_k^{H_۰})\\
&\qquad=\sup_{\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i^{H_۰},i=۱,\dots k}\prod_{i=۱}^{k}\prod_{j=۱}^{n_i} f_{\boldsymbol{X}_{ij}}(\boldsymbol{x}_{ij},\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i^{H_۰})\\
&\qquad=\prod_{i=۱}^{k}\prod_{r=۱}^{k_i}\sup_{\boldsymbol{\mu}_r^{(i)},\boldsymbol{\Sigma}_{rr}^{(i)},i=۱,\dots k}\prod_{j=۱}^{n_i} f_{\boldsymbol{X}_{rj}^{(i)}}(\boldsymbol{x}_{rj}^{(i)};\boldsymbol{\mu}_r^{(i)},\boldsymbol{\Sigma}_{rr}^{(i)})\\
&\qquad=\prod_{i=۱}^{k}\prod_{r=۱}^{k_i}\left(۲\pi e n_i^{-۱}\right)^{-\frac{n_i p_r^{(i)}}{۲}}\left|\boldsymbol{W}_{rr}^{(i)}\right|^{-\frac{n_i}{۲}}.
\end{aligned}$$

در نتیجه، آماره LRT عبارت خواهد بود از:

$$\begin{aligned}
W_{\boldsymbol{n}} &= \frac{\sup_{\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i^{H_۰},i=۱,\dots k} L(\boldsymbol{\mu}_۱,\dots,\boldsymbol{\mu}_k;\boldsymbol{\Sigma}_۱^{H_۰},\dots,\boldsymbol{\Sigma}_k^{H_۰})}{\sup_{\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,i=۱,\dots k} L(\boldsymbol{\mu}_۱,\dots,\boldsymbol{\mu}_k;\boldsymbol{\Sigma}_۱,\dots,\boldsymbol{\Sigma}_k)}\\
&= \frac{\prod_{i=۱}^{k}\prod_{r=۱}^{k_i}\left(\frac{۲\pi e}{n_i}\right)^{-\frac{n_i p_r^{(i)}}{۲}}\left|\boldsymbol{W}_{rr}^{(i)}\right|^{-\frac{n_i}{۲}}}{\prod_{i=۱}^{k}\left(\frac{۲\pi e}{n_i}\right)^{-\frac{n_i p_i}{۲}}|\boldsymbol{W}_i|^{-\frac{n_i}{۲}}}\\
&= \prod_{i=۱}^{k}\left(\frac{|\boldsymbol{W}_i|}{\prod_{r=۱}^{k_i}|\boldsymbol{W}_{rr}^{(i)}|}\right)^{\frac{n_i}{۲}},
\end{aligned}\tag{۲}$$

که در آن $\boldsymbol{n}=(n_۱,\dots,n_k)$. توجه شود که آماره $W_{\boldsymbol{n}}$ تنها در حالت $\min_{۱\le i\le k} n_i > p$ تعریف می‌شود. نظریه عام آزمون‌های نسبت درست‌نمایی (رائو، ۲۰۰۹) بیان می‌کند که وقتی اندازه‌های نمونه‌ای n_i، $i=۱,\dots,k$، از بُعد p بزرگ‌اند و نسبت‌های $\frac{p}{n_i}$ اعدادی نزدیک به صفر ($\frac{p}{n_i}\to ۰$) هستند، توزیع تحت فرض صفر آماره $-۲\log W_{\boldsymbol{n}}$ به توزیع $\chi^۲$ با $df=\frac{۱}{۲}(kp^۲-\sum_{i=۱}^{k}\sum_{r=۱}^{k_i}{p_r^{(i)}}^۲)$ درجه آزادی همگراست.

به هر حال، برای داده‌های با بُعد (نسبتاً) بالا، تقریب $\chi^۲$ عملاً غیر قابل استفاده است. به منظور غلبه بر این مشکل، در قالب قضیه زیر تقریب بهتری از تقریب خی‌دو برای توزیع تحت فرض صفر آماره LRT ارائه می‌شود. در ادامه، به منظور تسهیل نمایش و کار با روابط حدی فرض شده است که n_i به ازای هر $i=۱,\dots,k$ به صورت $n_i=n_i(p)$ به p وابسته است.

قضیه ۱. فرض کنید به ازای هر $i = ۱, \dots, k$، رابطه

$$n_i = n_i(p) > ۱ + p = ۱ + \sum_{j=۱}^{k_i} p_j^{(i)}$$

برقرار باشد، که در آن $(p_۱^{(i)}, , \dots, p_{k_i}^{(i)})$ افرازی از p است و به ازای هر $r = ۱, \dots, k_i$،

$$\lim_{p\to\infty} \frac{p_r^{(i)}}{n_i} = y_r^{(i)} \in (۰, ۱).$$

آنگاه تحت فرض $H_۰$ در (۱)، وقتی $p \longrightarrow \infty$ آماره $\frac{\log W_{\boldsymbol{n}} - \mu_{\boldsymbol{n}}}{n\sigma_{\boldsymbol{n}}}$ در توزیع به $N(۰, ۱)$ همگراست، به گونه‌ای که روابط

$$\mu_{\boldsymbol{n}} = \frac{۱}{۲}\sum_{i=۱}^{k}\left[\left(\sum_{r=۱}^{k_i} r^{۲}_{n_i-۱,p_r^{(i)}}(p_r^{(i)} - n_i + ۱/۵) - r^{۲}_{n_i-۱,p}(p - n_i + ۱/۵)\right)n_i\right]$$

$$\sigma^{۲}_{\boldsymbol{n}} = \frac{۱}{۲}\sum_{i=۱}^{k}\left[\left(r^{۲}_{n_i-۱,p} - \sum_{r=۱}^{k_i} r^{۲}_{n_i-۱,p_r^{(i)}}\right)\left(\frac{n_i}{n}\right)^{۲}\right]$$

برقرارند، که در آنها $n = \sum_{i=۱}^{k} n_i$ و $r_{x,y} = \sqrt{-\log(۱ - \frac{y}{x})}$.

برهان. هاردی و همکاران (۱۹۸۸) برای اعداد حقیقی $a_۱, , \cdots, a_q$ بزرگتر از -۱، که همگی یا مثبت‌اند یا منفی، ثابت کردند $\prod_{i=۱}^{q}(۱ + a_i) > ۱ + \sum_{i=۱}^{q} a_i$ یا

$$\log\left(\prod_{i=۱}^{q}(۱ + a_i)\right) - \log\left(۱ + \sum_{i=۱}^{q} a_i\right) > ۰.$$

با تثبیت $i \in \{۱, \dots, k\}$ و تعریف $a_i = -\frac{p_r^{(i)}}{n_i - ۱}$ و $q = k_i$، مشاهده می‌شود

$$r^{۲}_{n_i-۱,p} - \sum_{r=۱}^{k_i} r^{۲}_{n_i-۱,p_r^{(i)}} = \log\left(\prod_{r=۱}^{k_i}(۱ - \frac{p_r^{(i)}}{n_i - ۱})\right) - \log\left(۱ - \sum_{r=۱}^{k_i}\frac{p_r^{(i)}}{n_i - ۱}\right) > ۰.$$

در نتیجه، $\sigma^{۲}_{\boldsymbol{n}} > ۰$. از این وقتی $p \longrightarrow \infty$:

$$\frac{n}{p} = \sum_{i=۱}^{k}\frac{n_i}{p} = \sum_{i=۱}^{k}\left(\sum_{r=۱}^{k_i}\frac{p_r^{(i)}}{n_i}\right)^{-۱} \to \sum_{i=۱}^{k}\left(\sum_{r=۱}^{k_i} y_r^{(i)}\right)^{-۱} = \sum_{i=۱}^{k}\frac{۱}{y_i} := \frac{۱}{y},$$

که در آن $y_i = \sum_{r=۱}^{k_i} y_r^{(i)} \in (۰, ۱]$ و $y \in (۰, \frac{۱}{k}]$. در نتیجه، $\frac{n_i}{n} = \frac{\left(\frac{n_i}{p}\right)}{\left(\frac{n}{p}\right)} \to \frac{y}{y_i} \in (۰, ۱]$ وقتی $p \longrightarrow \infty$. بنابراین، $\sigma^{۲} = \lim_{p\to\infty}\sigma^{۲}_{\boldsymbol{n}}$ به ازای $\max_{۱\le i\le k} y_i < ۱$ برابر

$$\frac{۱}{۲}\sum_{i=۱}^{k}\left[\left(\sum_{r=۱}^{k_i}\log(۱ - y_r^{(i)}) - \log(۱ - y_i)\right)\left(\frac{y}{y_i}\right)^{۲}\right]$$

و به ازای $\max_{۱\le i\le k} y_i = ۱$ برابر $+\infty$ خواهد بود. در حقیقت برای حالت دوم، از اینکه $y_r^{(i)} \in (۰, ۱)$ و $\lim_{x\to ۱^-} \log(۱-x) = -\infty$ واضح است که حد برابر $+\infty$ خواهد شد. حال با تثبیت s با شرط $|s| < \frac{\sigma}{۲y}$، t به شکل $t = t_{\boldsymbol{n}} = \frac{s}{n\sigma_{\boldsymbol{n}}}$ تعریف می‌شود. واضح است که $s > -\frac{n\sigma_n}{۲(p+۱)} \to -\frac{\sigma}{۲y}$ وقتی $p \longrightarrow \infty$. این موضوع نتیجه می‌دهد که برای p به اندازه کافی بزرگ $s > -\frac{n\sigma_n}{۲(p+۱)}$ یا $\frac{s}{n\sigma_n} > -\frac{۱}{۲(p+۱)}$. این نابرابری در کنار این واقعیت که

$$\max_{۱\le i\le k}\Big\{\frac{۱}{n_i}\Big\} < \frac{۱}{p+۱} \quad \text{یا} \quad \max_{۱\le i\le k}\Big\{\frac{p}{n_i} - ۱\Big\} < -\frac{۱}{p+۱} < -\frac{۱}{۲(p+۱)},$$

نتیجه می‌دهد که $t = t_{\boldsymbol{n}} = \frac{s}{n\sigma_{\boldsymbol{n}}} > \max_{۱\le i\le k}\Big\{\frac{p}{n_i} - ۱\Big\}$ وقتی $p \longrightarrow \infty$. حال برای یک مقدار ثابت i و حالتی که $y_i < ۱$، $\frac{r^۲_{n_i-۱,p}}{\sigma_n^۲}$ برای $\max_{۱\le i\le k} y_i < ۱$ به مقدار $\frac{-\log(۱-y_i)}{\sigma^۲}$ و به ازای $\max_{۱\le i\le k} y_i = ۱$ به مقدار ۰ همگرا است. این نتیجه می‌دهد که $\frac{r_{n_i-۱,p}}{\sigma_n}$ به ازای $\max_{۱\le i\le k} y_i < ۱$ به $\sqrt{\frac{-\log(۱-y_i)}{\sigma^۲}}$ و به ازای $\max_{۱\le i\le k} y_i = ۱$ به ۰ همگرا خواهد بود. علاوه بر این، برای حالت $y_i = ۱$،

$$۲\sigma_n^۲ \ge \big(r^۲_{n_i-۱,p} - \sum_{r=۱}^{k_i} r^۲_{n_i-۱,p_r^{(i)}}\big)\Big(\frac{n_i}{n}\Big)^۲,$$

یا به طور معادل

$$\frac{r_{n_i-۱,p}}{\sigma_n} \le \sqrt{۲\Big(\frac{n_i}{n}\Big)^۲ + \sigma_n^{-۲}\sum_{r=۱}^{k_i} r^۲_{n_i-۱,p_r^{(i)}}} \to \frac{\sqrt{۲}}{y} < \infty,$$

وقتی $p \longrightarrow \infty$. بنابراین،

$$\limsup_{p\to\infty} \frac{r_{n_i-۱,p}}{\sigma_n} < \infty \Rightarrow \frac{۱}{\sigma_n} = O(\frac{۱}{r_{n_i-۱,p}}),$$

یا

$$\frac{tn_i}{۲} = \frac{s}{۲}\frac{n_i}{n}\frac{۱}{\sigma_n} = O(۱)\frac{۱}{\sigma_n} = O(\frac{۱}{r_{n_i-۱,p}}).$$

به طور مشابه، از آنجا که $-\log(۱-x)$ برای $x < ۱$ یک تابع صعودی است،

$$p_r^{(i)} < p \Leftrightarrow r^۲_{n_i-۱,p_r^{(i)}} < r^۲_{n_i-۱,p}.$$

در نتیجه،

$$\limsup_{p\to\infty} \frac{r_{n_i-۱,p_r^{(i)}}}{\sigma_n} < \limsup_{p\to\infty} \frac{r_{n_i-۱,p}}{\sigma_n} < \infty,$$

یا

$$\frac{۱}{\sigma_n} = O(\frac{۱}{r_{n_i-۱,p_r^{(i)}}}) \Rightarrow \frac{tn_i}{۲} = O(\frac{۱}{r_{n_i-۱,p}}).$$

حال بنا بر قضیه ۱۱.۲.۳ در مؤیرهد (۱۹۸۲)، وقتی $H_۰$ صحیح است، tامین گشتاور $W_{\boldsymbol{n}}$ برابر

$$\begin{aligned} E\left[W_{\boldsymbol{n}}^t\right] &= \prod_{i=۱}^{k} E\left[(|\boldsymbol{W}_i| \prod_{r=۱}^{k_i} \left|\boldsymbol{W}_{rr}^{(i)}\right|^{-۱})^{\frac{n_i t}{۲}}\right] \\ &= \prod_{i=۱}^{k} \left(\frac{\Gamma_{p_i}\left(\frac{n_i-۱}{۲} + \frac{n_i t}{۲}\right)}{\Gamma_{p_i}\left(\frac{n_i-۱}{۲}\right)} \prod_{r=۱}^{k_i} \frac{\Gamma_{p_r^{(i)}}\left(\frac{n_i-۱}{۲}\right)}{\Gamma_{p_r^{(i)}}\left(\frac{n_i-۱}{۲} + \frac{n_i t}{۲}\right)} \right), \end{aligned} \tag{۳}$$

خواهد بود، که در آن برای عدد مختلط z با ویژگی $Re(z) > \frac{۱}{۲}(p-۱)$، تابع گامای چند متغیره[11] $\Gamma_p(z)$ (مؤیرهد، ۱۹۸۲، صفحه ۶۲) به صورت $\Gamma_p(z) = \pi^{\frac{p(p-۱)}{۴}} \prod_{j=۱}^{p} \Gamma\left(z - \frac{۱}{۲}(j-۱)\right)$ تعریف می‌شود. بدیهی است که امید ریاضی (۳) تنها زمانی وجود دارد که $\frac{n_i-۱}{۲} + \frac{tn_i}{۲} > \frac{p-۱}{۲}$، $i \in \{۱, ۲, \ldots, k\}$ یا $t > \max_{۱\le i\le k}\left\{\frac{p}{n_i} - ۱\right\}$. با توجه به اینکه

$$t = t_{\boldsymbol{n}} = \frac{s}{n\sigma_{\boldsymbol{n}}} > \max_{۱\le i\le k}\left\{\frac{p}{n_i} - ۱\right\}, \quad \frac{tn_i}{۲} = O(\frac{۱}{r_{n_i-۱,p}}),$$

و $\frac{tn_i}{۲} = O(\frac{۱}{r_{n_i-۱,p_r^{(i)}}})$ وقتی که $p \longrightarrow \infty$، با استفاده از لم ۵.۴ ژیانگ و یانگ (۲۰۱۳)، به برابری‌های

$$\begin{aligned} \log \frac{\Gamma_p\left(\frac{n_i-۱}{۲} + \frac{tn_i}{۲}\right)}{\Gamma_p\left(\frac{n_i-۱}{۲}\right)} &= \frac{tpn_i}{۲} \log\left(\frac{n_i - ۱}{۲e}\right) + r_{n_i-۱,p}^{۲} \frac{n_i^{۲} t^{۲}}{۴} \\ &\quad - r_{n_i-۱,p}^{۲} \left(p - n_i + ۱/۵\right) \frac{tn_i}{۲} + o(۱), \end{aligned}$$

و

$$\begin{aligned} \log \frac{\Gamma_{p_r^{(i)}}\left(\frac{n_i-۱}{۲} + \frac{tn_i}{۲}\right)}{\Gamma_{p_r^{(i)}}\left(\frac{n_i-۱}{۲}\right)} &= \frac{tp_r^{(i)} n_i}{۲} \log\left(\frac{n_i - ۱}{۲e}\right) + r_{n_i-۱,p_r^{(i)}}^{۲} \frac{n_i^{۲} t^{۲}}{۴} \\ &\quad - r_{n_i-۱,p_r^{(i)}}^{۲} \left(p_r^{(i)} - n_i + ۱/۵\right) \frac{tn_i}{۲} + o(۱), \end{aligned}$$

[11] Multivariate gamma function

حاصل می‌شود. بنابراین،

$$\begin{aligned}\log E\big[W_{\boldsymbol{n}}^{t}\big] &= \sum_{i=۱}^{k}\Big[\log\frac{\Gamma_p\big(\frac{n_i-۱}{۲}+\frac{tn_i}{۲}\big)}{\Gamma_p\big(\frac{n_i-۱}{۲}\big)} - \sum_{r=۱}^{k_i}\log\frac{\Gamma_{p_r^{(i)}}\big(\frac{n_i-۱}{۲}+\frac{tn_i}{۲}\big)}{\Gamma_{p_r^{(i)}}\big(\frac{n_i-۱}{۲}\big)}\Big]\\ &= \sum_{i=۱}^{k}\Big[\frac{tpn_i}{۲}\log\big(\frac{n_i-۱}{۲e}\big)+r^{۲}_{n_i-۱,p}\frac{n_i^{۲}t^{۲}}{۴}\Big]\\ &\quad-\sum_{i=۱}^{k}\Big[r^{۲}_{n_i-۱,p}\big(p-n_i+۱/۵\big)\frac{tn_i}{۲}+o(۱)\Big]\\ &\quad-\sum_{i=۱}^{k}\sum_{r=۱}^{k_i}\Big[\frac{tp_r^{(i)}n_i}{۲}\log\big(\frac{n_i-۱}{۲e}\big)+r^{۲}_{n_i-۱,p_r^{(i)}}\frac{n_i^{۲}t^{۲}}{۴}\Big]\\ &\quad+\sum_{i=۱}^{k}\sum_{r=۱}^{k_i}\Big[r^{۲}_{n_i-۱,p_r^{(i)}}\big(p_r^{(i)}-n_i+۱/۵\big)\frac{tn_i}{۲}+o(۱)\Big]\\ &= \frac{۱}{۲}n^{۲}\sigma_{\boldsymbol{n}}^{۲}t^{۲}+\mu_{\boldsymbol{n}}t+o(۱),\end{aligned}$$

وقتی $p\longrightarrow\infty$. از آنجا که $t=t_{\boldsymbol{n}}=\frac{s}{n\sigma_{\boldsymbol{n}}}$،

$$\log E\big[e^{\frac{\log W_{\boldsymbol{n}}}{n\sigma_n}s}\big]=\log e^{\frac{۱}{۲}s^{۲}+\frac{\mu_{\boldsymbol{n}}}{n\sigma_n}s+o(۱)},$$

یا به طور معادل

$$E\big[e^{\frac{\log W_{\boldsymbol{n}}-\mu_n}{n\sigma_n}s}\big]=e^{\frac{۱}{۲}s^{۲}+o(۱)},$$

وقتی $p\longrightarrow\infty$. یعنی، وقتی $p\longrightarrow\infty$، $\frac{\log W_{\boldsymbol{n}}-\mu_n}{n\sigma_{\boldsymbol{n}}}$ در توزیع به $N(۰,۱)$ همگراست.

به طور خلاصه، قضیه فوق بیان می‌کند که در حالت داده‌های با بُعد نسبتاً بالا، توزیع تقریبی مناسب برای لگاریتم آماره نسبت درست‌نمایی $-۲\log W_{\boldsymbol{n}}$ توزیع نرمال با میانگین $-۲\mu_{\boldsymbol{n}}$ و واریانس $۴n^{۲}\sigma_{\boldsymbol{n}}^{۲}$ است. در بخش بعد، به مقایسه این تقریب با تقریب کلاسیک خی‌دو برای $-۲\log W_{\boldsymbol{n}}$ پرداخته می‌شود.

۳ مطالعه شبیه‌سازی

در این بخش به منظور مقایسه اندازه و توان آزمون‌های LRT برای فرض (۱) بر اساس توزیع‌های تقریبی خی‌دو ($\chi^{۲}$) و نرمال (N) یک مطالعه شبیه‌سازی انجام شده است. در این شبیه‌سازی‌ها،

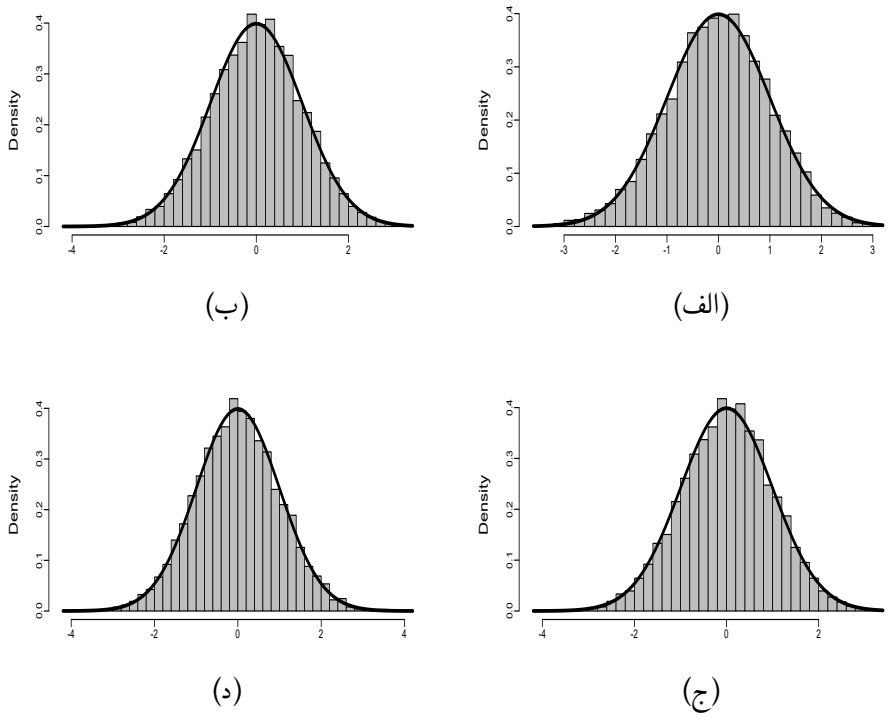
ماتریس‌های $p \times p$ ی $\boldsymbol{J}_p$ و $\boldsymbol{I}_p$ به ترتیب به عنوان ماتریسی با تمامی درایه‌های برابر با عدد یک و ماتریس همانی تعریف شده است. همچنین فرض می‌شود که ماتریس‌های کوواریانس $\boldsymbol{\Sigma}_i$ها به ازای ثابت $\circ \leq \rho < ۱$، همگی برابر $\rho\boldsymbol{J}_p + (۱-\rho)\boldsymbol{I}_p$ هستند. در حقیقت در اینجا توان آزمون به عنوان تابعی از ρ بررسی شده است که در آن به ازای $\rho = \circ$ توزیع تحت فرض صفر و با دور شدن ρ از عدد $\circ$ توزیع‌های تحت فرض مقابل بدست می‌آیند. علاوه بر این، بردارهای میانگین $\boldsymbol{\mu}_i$ها همگی برابر بردار صفر، سطح معنی‌داری اسمی α برابر ۰/۰۵ و $k = ۳$ است. همچنین، در هر یک از این k جامعه نرمال، به منظور آزمون همزمان استقلال زیربردارها به تفکیک هر یک از این جوامع، افرازهای p یا همان ابعاد زیربردارها در تمامی k گروه یکسان در نظر گرفته شده است. تمامی برنامه‌های کامپیوتری در نرم‌افزار آماری R نوشته شده و در صورت درخواست، از طرف نویسنده در اختیار خواننده قرار می‌گیرد.

حال برای هر ترکیب از پارامترهای شبیه‌سازی ρ، $n_۱$، $n_۲$، $n_۳$ و p که در جدول ۱ آمده‌اند، با استفاده از ۱۰۰۰۰ نمونه از توزیع $N_p(\circ, \rho\boldsymbol{J}_p + (۱-\rho)\boldsymbol{I}_p)$، مقادیر شبیه‌سازی‌شده اندازه آزمون $\hat{\varphi}_M(\circ)$ و توان آزمون $\hat{\varphi}_M(\rho)$، $\rho > \circ$ برای تقریب $M \in \{\chi^۲, N\}$ محاسبه و در جدول ۱ آورده شده است. برای نمونه‌های شبیه‌سازی‌شده تحت فرض صفر ($\rho = \circ$) و نمونه‌های شبیه‌سازی‌شده تحت فرض مقابل ($\rho > \circ$) به ترتیب مقادیر $\hat{\varphi}_M(\circ)$ و $\hat{\varphi}_M(\rho)$ برای $M \in \{\chi^۲, N\}$ برابر نسبت تعداد دفعاتی که در این ۱۰۰۰۰ بار فرض صفر رد شده، محاسبه شده است. همانطور که در جدول ۱، ملاحظه می‌شود تقریب نرمال معرفی‌شده و تقریب کلاسیک خی‌دو برای اندازه‌های نمونه‌ای n_iی بزرگ و بُعد pی کوچک رفتاری شبیه یکدیگر دارند. چنین وضعیتی را می‌توان در جدول ۱ به ازای $\boldsymbol{n} = ۶\boldsymbol{\nu} = (۱۵\circ, ۱۵\circ, ۱۵\circ)$، $p = ۵$ و $\rho = \circ, \circ/\circ۵, \circ/۶$ ملاحظه کرد. تحت فرض $H_\circ$ با افزایش p و نزدیک شدن آن به مقادیر اندازه‌های نمونه‌ای n_i (حالت داده‌های با بُعد نسبتاً بالا)، اندازه آزمون با تقریب $\chi^۲$؛ یعنی، $\hat{\varphi}_{\chi^۲}(\circ)$ به جای اینکه به مقدار اسمی $\alpha = \circ/\circ۵$ نزدیک باشد، مقادیری نزدیک به عدد یک قبول می‌کند (!) در حالی که اندازه آزمون روش معرفی‌شده؛ یعنی، $\hat{\varphi}_N(\circ)$، در حدود مقدار اسمی ۰/۰۵ است (جدول ۱ را به ازای $\rho = \circ$ و مقادیر بزرگ p مشاهده کنید). بنابراین تقریب خی‌دو در مقایسه با تقریب نرمال برای بُعد نسبتاً بالا عملاً غیر قابل استفاده است. نگاهی بر جدول ۱ نشان می‌دهد که به ازای انحرافات کوچک از فرض صفر به مانند $\rho = \circ/\circ۵$، توان آزمون

جدول ۱: اندازه آزمون $\hat{\varphi}_M(\circ)$ و توان آزمون $\hat{\varphi}_M(\rho)$ برای تقریب $M \in \{\chi^2, N\}$

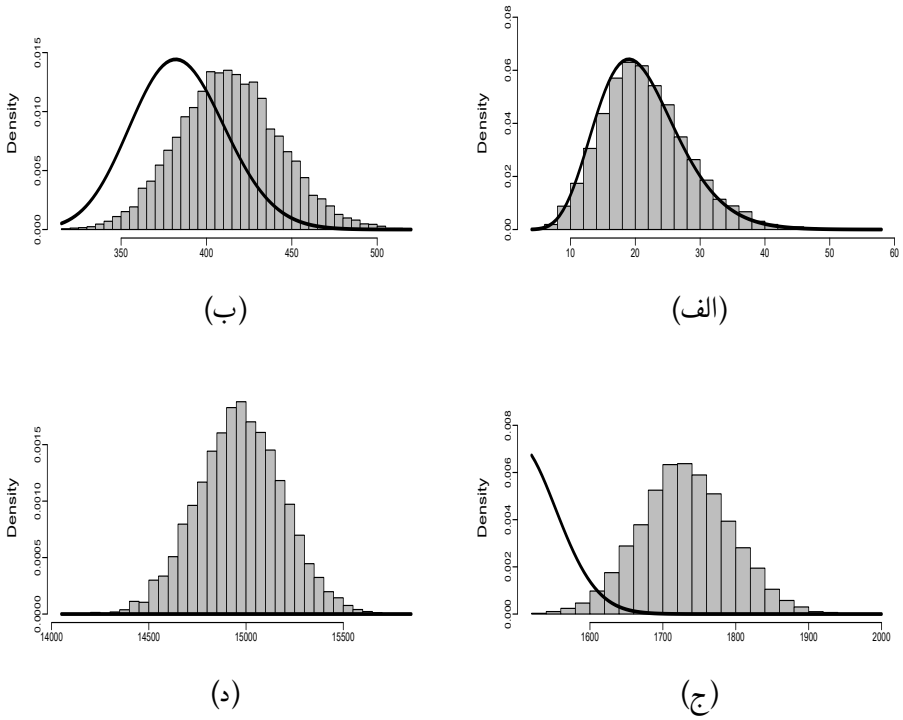
$\boldsymbol{n=\nu}$				$\boldsymbol{n=4\nu}$			
ρ	p	افراز p	$\hat{\varphi}_{\chi^2}(\rho)$	$\hat{\varphi}_N(\rho)$	p	افراز p	$\hat{\varphi}_{\chi^2}(\rho)$	$\hat{\varphi}_N(\rho)$
	۵	۲, ۱, ۲	۰/۱۵۴	۰/۰۴۸	۵	۳, ۲	۰/۰۶۶	۰/۰۴۶
۰	۱۰	۴, ۳, ۳	۰/۶۶۱	۰/۰۵۲	۵۰	۱۵, ۱۷, ۱۸	۱	۰/۰۵
	۲۰	۴, ۵, ۶, ۵	۱	۰/۰۶۲	۹۵	۱۵, ۳۰, ۳۰, ۲۰	۱	۰/۰۵۵
	۵	۲, ۱, ۲	۰/۱۹۶	۰/۰۶۸	۵	۳, ۲	۰/۱۹۳	۰/۱۳۸
۰/۰۵	۱۰	۴, ۳, ۳	۰/۷۳۹	۰/۰۷۰	۵۰	۱۵, ۱۷, ۱۸	۱	۰/۴۷۴
	۲۰	۴, ۵, ۶, ۵	۱	۰/۰۶۸	۹۵	۱۵, ۳۰, ۳۰, ۲۰	۱	۰/۲۲۳
	۵	۲, ۱, ۲	۱	۱	۵	۳, ۲	۱	۱
۰/۶	۱۰	۴, ۳, ۳	۱	۱	۵۰	۱۷, ۱۸, ۱۵	۱	۱
	۲۰	۴, ۵, ۶, ۵	۱	۱	۹۵	۱۵, ۳۰, ۳۰, ۲۰	۱	۱
$\boldsymbol{n=2\nu}$				$\boldsymbol{n=6\nu}$			
	۵	۲, ۳	۰/۰۸۹	۰/۰۴۷	۵	۱, ۴	۰/۰۵۷	۰/۰۴۲
۰	۲۰	۵, ۷, ۸	۰/۹۶۳	۰/۰۵۱	۸۰	۱۲, ۲۴, ۱۳, ۲۱, ۱۰	۱	۰/۰۵۴
	۴۵	۱۷, ۱۵, ۱۳	۱	۰/۰۶۱	۱۴۵	۵۰, ۷۰, ۲۵	۱	۰/۰۵۴
	۵	۲, ۳	۰/۱۴۹	۰/۰۷۹	۵	۱, ۴	۰/۲۱۴	۰/۱۷۴
۰/۰۵	۲۰	۵, ۷, ۸	۰/۹۹۴	۰/۱۲۷	۸۰	۱۲, ۲۴, ۱۳, ۲۱, ۱۰	۱	۰/۹۴۷
	۴۵	۱۷, ۱۵, ۱۳	۱	۰/۰۹۲	۱۴۵	۵۰, ۷۰, ۲۵	۱	۰/۲۲۴
	۵	۲, ۳	۱	۱	۵	۱, ۴	۱	۱
۰/۶	۲۰	۵, ۷, ۸	۱	۱	۸۰	۱۲, ۲۴, ۱۳, ۲۱, ۱۰	۱	۱
	۴۵	۱۷, ۱۵, ۱۳	۱	۱	۱۴۵	۵۰, ۷۰, ۲۵	۱	۱

با تقریب خی‌دوی کلاسیک به شکل کاذب بسیار بالاتر از توان آزمون پیشنهادی قرار می‌گیرد، که این به دلیل متورم یا بالا بودن خطای نوع اول آزمون با تقریب خی‌دو است. به هر حال با افزایش ρ به مقدار $\rho = ۰/۶$ و در نتیجه دور شدن بیشتر از فرض صفر استقلال، توان دو آزمون رفته رفته بر یکدیگر منطبق و به عدد یک همگرا می‌شوند. نتایج حاصل از جدول ۱ برای نمونه‌های تحت فرض $H_۰$ را می‌توان در شکل‌های ۱ و ۲ نیز مشاهده کرد.



شکل ۱: بافت‌نگار مقادیر شبیه‌سازی‌شده و چگالی $N(۰,۱)$ به ازای $k = ۳$، الف- بُعد ۵ و افراز $(۱,۱,۳)$؛ ب- بُعد ۲۰ و افراز $(۸,۸,۴)$؛ ج- بُعد ۴۰ و افراز $(۱۰,۱۰,۲۰)$؛ د- بُعد ۱۰۰ و افراز $(۳۰,۳۰,۴۰)$.

در این شکل‌ها، به ازای مقادیر مختلف از پارامترهای شبیه‌سازی، نمودارهای بافت‌نگار بدست آمده از ۱۰۰۰۰ مقدارشبیه‌سازی شده از $\frac{\log W_n - \mu_n}{n\sigma_n}$ و $-۲\log W_n$ به ترتیب به همراه منحنی‌های متناظر با چگالی‌های $N(۰,۱)$ و $\chi^۲_{df}$ رسم شده است. شکل ۱ نشان می‌دهد که بافت‌نگار شبیه‌سازی‌شده



شکل ۲: بافت‌نگار مقادیر شبیه‌سازی‌شده و چگالی χ^2_{df} به ازای $k = ۳$، الف- بُعد ۵ و افراز $(۱, ۱, ۳)$؛ ب- بُعد ۲۰ و افراز $(۸, ۸, ۴)$؛ ج- بُعد ۴۰ و افراز $(۱۰, ۱۰, ۲۰)$؛ د- بُعد ۱۰۰ و افراز $(۳۰, ۳۰, ۴۰)$.

$\frac{\log W_{\boldsymbol{n}} - \mu_{\boldsymbol{n}}}{n\sigma_{\boldsymbol{n}}}$ و تابع چگالی $N(۰, ۱)$ با افزایش p بر یکدیگر منطبق می‌شوند؛ این در حالی است که در شکل ۲ بافت‌نگار شبیه‌سازی‌شده $-۲\log W_{\boldsymbol{n}}$ با افزایش p به مرور از چگالی χ^2_{df} دور می‌شود. به عبارت دیگر، در حالتی که p متناسب با n_iها افزایش می‌یابد یا نسبت‌های $\frac{p}{n_i}$ها به یک نزدیک می‌شوند، تقریب χ^2 در مقایسه با تقریب نرمال بدتر می‌شود. جالب این است که حتی به ازای بُعد کوچک نیز تقریب نرمال عملکرد خوبی دارد.

۴ تحلیل داده‌های سرطان پروستات

در این بخش کاربردی از روش پیشنهادی بر مجموعه داده‌های سرطان پروستات یا "سطح بیان ژن تومور پروستات" [۱۲] (سینگ و همکاران ، ۲۰۰۲؛ دتلینگ و بوهمن، ۲۰۰۲) ارائه شده است. این مجموعه از داده‌ها که در بسته spls [۱۳] از نرم‌افزار R نیز قابل دسترس است، شامل ۵۲ $= n_1$ نمونه از بافت‌های تومور پروستات و ۵۰ $= n_2$ نمونه ار بافت‌های سالم است. بر روی این دو جامعه از بافت‌ها، برای هر نمونه سطح بیان ۶۰۳۳ ژن به کمک تکنولوژی اَفی‌متریکس[۱۴] اندازه‌گیری و ثبت شده است. سینگ و همکاران (۲۰۰۲) بر اساس همبستگی‌های بین مقادیر سطوح بیان و همچنین همبستگی بین این سطوح با متغیرهای بالینی و آسیب‌شناختی (به مانند سن بیمار، نتیجه آزمون‌های تشخیصی PSA[۱۵] و GS[۱۶]، غیره) به تشریح رفتار بالینی سرطان پرستات پرداختند.

به دلیل اهمیت همبستگی‌های درونی موجود در هر یک از این دو جامعه، در سطح آزمون $\alpha =$ ۰٫۰۵ استقلال همزمان زیربردارهایی از $p =$ ۴۸ سطح بیان ژن نخست از بین ۶۰۳۳ سطح بیان مطالعه می‌شود. در این مطالعه، به عنوان یک حالت خاص از فرض (۱)، استقلال تمامی ۴۸ سطح بیان ژن یا همان آزمون کُرویت ژیانگ و یانگ (۲۰۱۳) به طور همزمان در هر دوجامعه بافت‌های سالم و بافت‌های سرطانی بررسی شده است. توجه شود که این حالت (به بخش ۲ مراجعه شود) افراز مشترکی از $p =$ ۴۸ سطح بیان در دو جامعه به صورت $p = (p_1, \ldots, p_{48}) = (1, \ldots, 1)$ را نتیجه می‌دهد. همانگونه که مشاهده می‌شود در این حالت، نسبت‌های $\frac{p}{n_1} = \frac{48}{52} =$ ۰٫۹۲۳ و $\frac{p}{n_2} = \frac{48}{50} =$ ۰٫۹۶۰ حاصل می‌شوند که اعدادی نزدیک به یک‌اند (یعنی داده‌های با بُعد نسبتاً بالا) و در نتیجه روش LRT با تقریب خی‌دو به منظور آزمون همزمان استقلال زیربردارهای در هر دو جامعه بافت‌های سالم و بافت‌های سرطانی عملاً غیر قابل استفاده است. نتایج حاصل از آزمون پیشنهادی و روش کلاسیک LRT به شرح زیر است. مقادیر آماره‌های آزمون LRT با تقریب خی‌دو و تقریب پیشنهادی نرمال به ترتیب برابر $-2\log(W_n) =$ ۱۱۳۵۴٫۴۱ و $\frac{\log W_n - \mu_n}{n\sigma_n} =$ −۴۴٫۶۹۲ ($\mu_n =$ −۲۱۶۵٫۸۲ و $\sigma_n =$ ۰٫۵۹۳) بدست می‌آیند که به ترتیب در مقایسه با مقادیر بحرانی

[12] Prostate tumor gene expression level dataset

[13] https://CRAN.R-project.org/package=spls

[14] Affymetrix technology

[15] Prostate-specific antigen test

[16] Gleason score

$\chi^2_{df}(۰/۹۵) = \chi^2_{۲۲۵۶}(۰/۹۵) = ۲۳۶۷/۶۱۲$ و $z_{\frac{۰۵}{۲}} = ۱/۶۴۵$، رد فرض استقلال سطح بیان هم در جامعه بافت‌های سالم و هم در جامعه بافت‌های سرطانی را نتیجه می‌دهد. همانطور که ملاحظه می‌شود، روش کلاسیک LRT در مقایسه با روش پیشنهادی این مقاله با شدت بالاتری فرض صفر را رد می‌کند که این به دلیل متورم بودن خطای نوع اول این آزمون برای داده‌هائی نسبتاً با بعد بالا است؛ که می‌توانست حتی در صورت مستقل بودن مؤلفه‌ها نیز رأی به عدم استقلال مؤلفه‌ها دهد! به عبارتی دیگر، این مثال بیان می‌کند که در حالت داده‌های با بعُد نسبتاً بالا (حالت‌هایی که نسبت بعُد به اندازه نمونه کمتر از یک و نزدیک یک‌اند) باید در استفاده از روش LRT با تقریب کلاسیک خی‌دو تجدید نظر نمود.

۵ بحث و نتیجه‌گیری

در این مقاله، آزمون همزمان استقلال زیربردارهای k بردار نرمال $p-$متغیره با بُعد نسبتاً بالا مورد مطالعه قرار گرفت. در حالتی که n_iها از بُعد p بزرگ‌اند و نسبت $\frac{p}{n_i}$ عددی کوچک و نزدیک به صفر است، این آزمون را می‌توان به کمک روش نسبت‌درست‌نمایی با تقریب قابل قبول خی‌دو انجام داد. اما برای داده‌های با بُعد نسبتاً بالا؛ یعنی، حالتی که در آن n_iها از بُعد p بزرگ‌اند و نسبت‌های $\frac{p}{n_i}$ اعدادی نزدیک به یک هستند، روش آزمون کلاسیک LRT با تقریب خی‌دو به جای اینکه به مقدار اسمی $\alpha = ۰/۰۵$ نزدیک باشد، مقادیری نزدیک به عدد یک قبول می‌کند که این خود به معنای غیر قابل استفاده بودن این آزمون برای داده‌های با بُعد نسبتاً بالا است. برای این حالت، در این مقاله نشان داده شد که تقریب مناسب‌تر برای توزیع تحت فرض صفر آماره LRT به جای تقریب خی‌دو تقریب نرمال است. نتایج شبیه‌سازی حاکی از این بود که برای داده‌های با بُعد نسبتاً بالا تقریب خی‌دو کارایی نداشته و تقریب نرمال جایگزین مناسبتری برای آن است. از این رو، به منظور بررسی همزمان استقلال زیربردارهای چندین بردار نرمال $p-$متغیره، که استقلال همزمان مؤلفه‌های این بردارها حالت خاصی از آن است، برای داده‌های با بُعد نسبتاً بالا، استفاده از روش پیشنهادی این مقاله توصیه می‌شود. توجه شود که برای داده‌های با بُعد بالا که در آن اندازه‌های نمونه‌ای از بُعد کوچکتراند، روش این مقاله کارساز نیست و نویسنده در حال تحقیق بر روی این حالت است.

تقدیر و تشکر

نویسنده مقاله از دو داور محترم، سردبیر و ویراستار مجله به دلیل تلاش ایشان در راستای ارتقای کیفی مقاله و برطرف نمودن ایرادات احتمالی و ارائه بهتر مقاله کمال تشکر و قدردانی را دارد.

مراجع

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis (3nd ed.)*. John Wiley & Sons, New York.

Bai, Z., D., Jiang, J., Yao, F. and Zheng. S. (2009), Corrections to LRT on Large-Dimensional Covariance Matrix by RMT, *The Annals of Statistics*, **37**, 3822-3840.

Chen, X. and Liu, W. (2018), Testing Independence with High-Dimensional Correlated Samples, *The Annals of Statistics*, **46**, 866-894.

Chen, S. X., Zhang, L. X. and Zhong, P. S. (2010), Tests for High-Dimensional Covariance Matrices, *Journal of the American Statistical Association*, **105**, 810-819.

Dettling, M. and Bühlmann, P. (2002), Supervised Clustering of Genes, *Genome biology* , **3**, Research0069-1.

Hardy, G., Littlewood, J., and Pólya, G. (1988), Inequalities, Reprint of the 1952 Edition, Cambridge Mathematical Library.

Jiang, D., Jiang, T. and Yang, F. (2012), Likelihood Ratio Tests for Covariance Matrices of High-Dimensional Normal Distributions, *Journal of Statistical Planning and Inference*, **142**, 2241-2256.

Jiang, T. and Yang, F. (2013), Central Limit Theorems for Classical Likelihood Ratio

Tests for High-Dimensional Normal Distributions, *The Annals of Statistics*, **41**, 2029-2074.

Ledoit, O. and Wolf, M. (2002), Some Hypothesis Tests for the Covariance Matrix when the Dimension is Large Compared to the Sample Size, *The Annals of Statistics*, **30**, 1081-1102.

Leung, D., and Drton, M. (2018), Testing Independence in High-Dimensions with Sums of Rank Correlations, *The Annals of Statistics*, **46**, 280-307.

Mao, G. (2018), Testing Independence in High-Dimensions using kendall's tau و *Computational Statistics & Data Analysis*, *117*, 128-137.

Muirhead, R. J. (1982), *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York.

Rao, C. (2009), *Linear Statistical Inference and its Applications*, Wiley Series in Probability and Statistics. Wiley.

Schott, J. R. (2005), Testing for Complete Independence in High-Dimensions, *Biometrika*, **92**, 951-956.

Schott, J. R. (2007), A Test for the Equality of Covariance Matrices when the Dimension is Large Relative to the Sample Sizes, *Computational Statistics & Data Analysis*, **51**, 6535-6542.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. and Lander, E. S. (2007), Gene Expression Correlates of Clinical Prostate Cancer Behavior, *Cancer cell* , **1**, 203-209.

Wilks, S. (1938), The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, *The Annals of Mathematical Statistics* , **9**, 60-62.

Simultaneous Test for Independence Among Subvectors of Several Moderately High Dimensional Multivariate Normal Distributions

Najarzadeh, D.

Department of Statistics, University of Tabriz, Tabriz, Iran.

Abstract: Testing the Hypothesis of independence of a p-variate vector subvectors, as a pretest for many others related tests, is always as a matter of interest. When the sample size n is much larger than the dimension p, the likelihood ratio test (LRT) with chisquare approximation, has an acceptable performance. However, for moderately high-dimensional data by which n is not much larger than p, the chisquare approximation for null distribution of the LRT statistic is no more usable. As a general case, here, a simultaneous subvectors independence testing procedure in all k p-variate normal distributions is considered. To test this hypothesis, a normal approximation for the null distribution of the LRT statistic was proposed. A simulation study was performed to show that the proposed normal approximation outperforms the chisquare approximation. Finally, the proposed testing procedure was applied on prostate cancer data.

Keywords: Multivariate normal distribution, likelihood ratio test, high-dimensional data, testing independence, multivariate Gamma function.

Mathematics Subject Classification (2010): 62H15, 62H10, 60F05.