دانشکده علوم ریاضی
گروه آمار

پایان‌نامه

برای دریافت درجه کارشناسی ارشد در رشته
آمار، گرایش آماررياضی

عنوان

برآورد انقباضی در تحلیل خوشه‌ای $K-$میانگین در توزیع نرمال چندمتغیره

استاد راهنما

دکتر محمد آرشی

دانشجو

پیمان برآبادی

بهمن ماه ۱۳۹۳

تقدیم به
پدر و مادر عزیزم

حمد و سپاس خدایی را که لطف و کرمش شامل من شد و این توانایی را به من عطا نمود تا گامی کوچک در اقیانوس بی‌کران علم و معرفت بردارم.

اکنون که با لطف خداوند متعال، این دوره تحصیلی را به پایان رسانده‌ام، ابتدا، از زحمات بی‌دریغ استاد راهنمای محترم، جناب آقای دکتر محمد آرشی که از ابتدای راه و در طی انجام تحقیق، با راهنمایی‌های خود مرا در نگارش این اثر یاری نمودند، سپاسگزاری می‌کنم. بر خود لازم می‌دانم که از اساتید محترم داور، جناب آقایان دکتر شاهسونی و دکتر باغیشنی، و همچنین از تمام اساتید فرهیخته گروه آمار دانشگاه شاهرود تشکر و قدردانی نمایم.

وظیفه خود می‌دانم که از زحمات پدر و مادر بسیار عزیز، دلسوز و فداکارم که پیوسته راهنمای من در زندگی بوده‌اند، قدردانی به عمل آورم. همچنین از برادرانم و خواهرم که همواره پشتیبان من بوده‌اند، صمیمانه سپاسگزاری می‌کنم.

با سپاس بی‌دریغ خدمت دوستان و همکلاسی‌های گران‌مایه‌ام، آقایان حمید کرمی کبیر، سید محمد حسینی، احسان اسحقی، میعاد ولیپور، علی یعقوبی، محسن مهدی‌زاده ، سعید افچنگی، محمد گواهی و خانم مینا نوروزی‌راد و آقای هادی خسروشاه که مرا صمیمانه و مشفقانه یاری داده‌اند.

در پایان از همه دوستان گرامی و کسانی که در گردآوری این مجموعه حامی و پشتیبان من بوده‌اند، کمال تشکر را دارم.

سیمان برآبادی
بهمن ماه ۱۳۹۳

تعهدنامه

اینجانب **پیمان برآبادی** دانشجوی کارشناسی ارشد رشته **آمار** **دانشکده علوم ریاضی** دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان **برآورد انقباضی در تحلیل خوشه‌ای $K-$میانگین در توزیع نرمال چندمتغیره** ، تحت راهنمایی **دکتر محمد آرشی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه شاهرود متعلق دارد، و مقالات مستخرج با نام " دانشگاه شاهرود " یا " Shahrood University " به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده) شده است، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

پیمان برآبادی
بهمن ماه ۱۳۹۳

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

در این پایان‌نامه یک الگوریتم عمومی برای بهبود دقت در تحلیل خوشه‌ای K-میانگین را که در آن از تأثیر برآوردگرهای انقباضی استفاده می‌شود، مطالعه می‌کنیم. مرکز خوشه‌ها را نسبت به میانگین کلی تمامی داده‌ها با استفاده از برآوردگرهای انقباضی به‌دست آورده و سپس برآوردگرهای انقباضی به عنوان مرکز خوشه جدید در تکرارهای خوشه‌بندی تا رسیدن به همگرایی استفاده می‌شود. نتایج به‌دست آمده از روش انقباضی را با روش متداول K-میانگین مقایسه می‌کنیم. اندازه‌گیری افزایش دقت خوشه‌بندی این روش‌ها با استفاده از شاخصی موسوم به رَند انجام می‌پذیرد. در این راستا می‌توان، با شبیه‌سازی، بهبود دقت خوشه‌بندی وابسته به دو عامل واریانس درون‌خوشه‌ای و بعد ماتریس کواریانس را بررسی کرده و ثابت کرد که دقت در داده‌های شبیه‌سازی‌شده و مثال‌هایی با داده‌های واقعی به طور معناداری افزایش می‌یابد.

کلمات کلیدی:

خوشه‌بندی، K-میانگین، برآوردگر انقباضی، برآوردگر بارانچیک، شاخص رَند.

پیش‌گفتار

تحلیل خوشه‌ای یک اصطلاح کلی برای حیطه‌ی وسیعی از روش‌های عددی می‌باشد که با هدف مشترک آشکار کردن یا پیدا کردن گروه‌ها یا خوشه‌هایی از مشاهدات هم‌جنس همراه می‌باشد که جدا از گروه‌های دیگر هستند. روش‌های خوشه‌بندی اصولا سعی دارند به طور قانون‌مند مشاهدات صورت گرفته توسط انسان‌ها را به‌درستی به چند گروه تقسیم کنند. روش خوشه‌بندی $K-$میانگین به‌عنوان یکی از مهمترین روش‌های خوشه‌بندی در علوم مختلف شناخته شده است. هدف ما بهبود این روش با استفاده از برآوردگرهای انقباضی می‌باشد که مخصوصا برای بعدهای بالا و واریانس‌های درون خوشه‌ای بزرگ کاربرد دارد. در سال (۲۰۰۹) برآوردگر جیمز-استاین برای بهبود روش $K-$میانگین سنتی معرفی شد. در این پایان‌نامه دو برآوردگر جیمز-استاین تعمیم‌یافته و برآوردگر بارانچیک برای بهبود روش $K-$میانگین به عنوان رهیافتی جدید مورد بررسی قرار خواهد گرفت. فصل‌های مختلف این پایان‌نامه شامل مطالب زیر می‌باشد:

- در فصل اول، تاریخچه‌ای مختصر از کارهای انجام‌شده در زمینه‌ی خوشه‌بندی، مفاهیم اولیه تحلیل خوشه‌ای، روش‌های مختلف خوشه‌بندی، برآوردگر انقباضی و شاخص رَند بیان می‌شود.

- در فصل دوم، ابتدا به برآوردگر جیمز-استاین و قضایای مربوط به آن پرداخته شده و سپس برآوردگر انقباضی جیمز-استاین برای بهبود روش $K-$میانگین در حالت‌های مختلف مطالعه می‌شود که در آن برآوردگر انقباضی به‌عنوان مرکز خوشه‌ی جدید جایگزین میانگین خوشه می‌شود و شاخص رَند برای اندازه‌گیری دقت این دو روش به‌کار برده می‌شود. همچنین در این فصل مجموعه داده‌ی ژن‌ها برای ارزیابی عملی روش جیمز-استاین طرح شده است.

- در فصل سوم، به معرفی دو برآوردگر انقباضی جیمز-استاین تعمیم‌یافته و بارانچیک می‌پردازیم و این دو برآوردگر به‌عنوان جایگزین برای برآوردگر جیمز-استاین معرفی می‌شوند و برای حالت‌های مختلف خوشه‌بندی مورد بررسی قرار می‌گیرند. در ادامه، مجموعه داده‌ی سفال رومانو-بریتیش براساس دو روش جدید تحلیل شده و عملکرد آن‌ها برای داده‌های واقعی مورد سنجش قرار می‌گیرد. در پایان این فصل نتایج به‌دست آمده از روش‌های مورد مطالعه را مورد تحلیل قرار

می‌دهیم. لازم به ذکر است که تمامی نتایج فصل سوم جدید بوده و از نگارنده‌ی این مجموعه می‌باشد.

- در این پایان‌نامه، پیوست آ شامل تعاریف و پیوست ب شامل کدهای نوشته‌شده در محیط R برای رسم شکل‌ها و به‌دست آوردن شاخص رَند برای مقایسه روش‌های بیان شده می‌باشد.

از این مجموعه، مقالات زیر مستخرج گردیده است:

۱. برآبادی، پ. آرشی، م. نوروزی‌راد، م. (۱۳۹۳) کاربرد برآوردگر بارانچیک در تحلیل خوشه‌ای $K-$میانگین، هفتمین کنفرانس بین‌المللی انجمن ایرانی تحقیق در عملیات، دانشگاه سمنان.

۲. برآبادی، پ. آرشی، م. خسروشاه، ه. (۱۳۹۳) برآوردگر انقباضی در تحلیل خوشه‌ای $K-$میانگین در توزیع نرمال چندمتغیره، نشریه‌ی اندیشه آماری، در حال بررسی.

فهرست نشانه‌ها و نمادها

$\mathbf{X}$ بردار تصادفی

X متغیر تصادفی

I_p ماتریس همانی $p \times p$

o برداری با درایه‌های صفر

Q ماتریس کواریانس

Q^{-1} معکوس ماتریس Q

Q' و Q^T ترانهاده ماتریس Q

$tr(Q)$ اثر ماتریس Q

$det(Q)$ دترمینان ماتریس Q

$\|X\|$ نرم ماتریس X

∂ عملگر مشتق

$R(\theta, X)$ تابع مخاطره‌ی پارامتر θ برای متغیر تصادفی X

$\lambda_{max}(Q)$ بزرگترین مقدار ویژه ماتریس Q

$N_p(\mu, \Sigma)$ نرمال p-متغیره با بردار میانگین μ و ماتریس کواریانس Σ

فهرست مطالب

لیست تصاویر

لیست جداول

فصل ۱

مقدمات و تعاریف

۱.۱ تاریخچه

خوشه‌بندی و طبقه‌بندی دارای سابقه طولانی هستند و نقش مهمی را در بسیاری از رشته‌های علمی ایفا می‌کنند. اکثر نرم‌افزارهای آماری دارای توضیحات ویژه و رویه‌های انجام تحلیل خوشه‌ای می‌باشند.
هدف از تشکیل خوشه‌ها آن است که در هر خوشه عضوهایی قرار دهیم که دارای واریانس کمتری نسبت به واریانس موجود بین خوشه‌ها باشند به طوری که اشیایی که دارای شباهت بیشتری با یکدیگر هستند در یک گروه قرار گیرند (فرشادفر، ۱۳۸۹).
رده‌های اصلی متعددی از روش‌های تحلیل خوشه‌ای ازجمله خوشه‌بندی سلسله مراتبی[۱] ، خوشه‌بندی افرازی[۲] و خوشه‌بندی مبتنی بر مدل[۳] وجود دارد. در این پایان‌نامه تاکید اصلی بر روی خوشه‌بندی K-میانگین است که رایج‌ترین روش افرازی می‌باشد. مک کوئین[۴] (۱۹۶۷) روش K-میانگین را به عنوان یک روش جایگزین برای روش خوشه‌بندی سلسله مراتبی معرفی کرد (برای مطالعه بیشتر به هارتیگان و وانگ[۵]، ۱۹۷۹ رجوع کنید). این روش به ویژه برای مجموعه داده‌های بزرگ و ابعاد بالا از روش‌های خوشه‌بندی سلسله مراتبی کارآمدتر است.

---

[۱] Hierarchical clustering
[۲] Partitional clustering
[۳] Model-based clustering
[۴] McQueen
[۵] Hartigan and Wong

تان[۶] و همکاران (۲۰۰۵) نشان دادند در روش $K-$میانگین، اگر تعداد خوشه‌ها (K) خیلی کوچک‌تر از تعداد مشاهدات (n) باشد، زمان محاسبات رابطه خطی با n دارد در حالی که زمان محاسبات خوشه‌بندی سلسله‌مراتبی به $n^۲$ بستگی دارد. این نتیجه‌گیری، روش K-میانگین را برای داده‌های بزرگ نسبت به روش سلسله مراتبی سودمندتر می‌کند. اجرای تحلیل خوشه‌ای K-میانگین به آسانی توسط اغلب نرم‌افزارهای آماری (مثلاً تابع $kmeans$ ارائه شده توسط گروه مرکزی توسعه‌ی R، ۲۰۰۹) قابل اجرا است. بسیاری از رهیافت‌های خوشه‌بندی که بر پایه اصل کمترین توان‌های دوم است، به‌دنبال قوی‌تر کردن خوشه‌بندی برای داده‌های پرت نسبت به الگوریتم K-میانگین می‌باشند. برای مثال، کوفمن و راسیو[۷] (۱۹۸۷) روش شناخته شده‌ی K-مدوید[۸] را به‌وسیله‌ی اجرای تابع pam در R گسترش دادند. کوئستا-آلبرتوس[۹] و همکاران (۱۹۹۷) K-میانگین تعدیل‌شده[۱۰] را معرفی کردند که در آن یک بخش خاص از اشیاء دورافتاده از خوشه‌بندی حذف شده است که این روش تعدیل کننده بعداً توسط کوئستا-آلبرتوس و همکاران (۲۰۰۸) و همچنین گارسیا-اسکودرو[۱۱] و همکاران (۲۰۰۹) توسعه یافت. علاوه‌براین بسیاری از محققین استفاده از ادغام روش K-میانگین با انتخاب متغیر را پیشنهاد کردند. برای مثال کزانوسکی و هاند[۱۲] (۲۰۰۹) یک روش برای تسریع انتخاب متغیرهای مفید قبل از خوشه‌بندی ارائه دادند. یکی دیگر از روش‌های انتخاب متغیر در این راستا توسط براسکو و کرادیت[۱۳] (۲۰۰۱) ارائه شد. وزن‌دهی به متغیر در خوشه‌بندی K-میانگین که توسط دیساربو[۱۴] و همکاران (۱۹۸۴) و ماکارنروف و لژندره[۱۵] (۲۰۰۱) پیشنهاد شد مزایایی مشابه با انتخاب متغیر داشت. استنلی[۱۶] در سال ۲۰۰۶ یک بررسی جامع و با ارزش از روش K-میانگین شامل خواص و بسیاری تغییرات آن انجام داد. استنلی و براسکو (۲۰۰۸a) یک روش انتخاب متغیر استانداردشده براساس

---

[۶] Tan
[۷] Kaufman and Rousseeuw
[۸] K-medoid
[۹] Cuesta-Albertos
[۱۰] Trimmed K-means
[۱۱] García-Escudero
[۱۲] Krzanowski and Hand
[۱۳] Brusco and Cradit
[۱۴] DeSarbo
[۱۵] Makarenkov and Legendre
[۱۶] Steinley

«نسبت واریانس به دامنه[۱۷]» معرفی کردند. برای مقایسه عملکرد انتخاب متغیر یا وزن‌دهی متغیرها به استنلی و براسکو (۲۰۰۸b) مراجعه کنید.

گائو و هیچکاک[۱۸] (۲۰۱۰) نشان دادند که برآوردگر انقباضی جیمز-استاین همان طور که باعث بهبود میانگین می‌گردد، می‌تواند باعث بهبود میانگین خوشه‌ها نیز شود و میانگین خوشه‌ها را به یکدیگر نزدیک‌تر کند.

یکی از ویژگی‌های روش جدید، در برگرفتن روش انقباضی است. این روش به‌طور کلی خوشه‌بندی داده‌های چندمتغیره با پراکندگی زیاد را در مقابل حالت‌های خاص با تعداد داده‌های پرت کنترل می‌کند. در شرایطی که تغییرات درون خوشه‌ای نسبتاً بالا است، این روش به‌طور معناداری خوشه‌بندی را نسبت به K-میانگین سنتی، به‌خصوص زمانی‌که بهبود خوشه‌بندی دشوار است، بهبود می‌بخشد.

۲.۱ مفاهیم اولیه تحلیل خوشه‌ای

۱.۲.۱ تعاریف پایه‌ای

در این بخش ابتدا چند تعریف را بیان می‌کنیم:

خوشه[۱۹]: اوریت و هوتورن[۲۰] (۱۹۷۴) چندین تعریف برای یک خوشه بیان کردند که در زیر می‌آوریم.

۱. هر خوشه مجموعه‌ای از نمونه‌های متشابه است و نمونه‌های خوشه‌های مختلف نامتشابه‌اند.

۲. خوشه، اجتماعی از نقاط در فضای نمونه است به طوری‌که فاصله‌ی بین هر دو نقطه در یک خوشه، از فاصله‌ی بین هر نقطه در یک خوشه و هر نقطه‌ای که در آن خوشه نیست کمتر است.

۳. خوشه را می‌توان ناحیه‌ای همبند از فضای چندبعدی شامل نقاطی با چگالی نسبتاً زیاد، جدا از ناحیه‌ای، شامل نقاط با چگالی کمتر توصیف کرد.

---

[۱۷] Variance-to-range ratio

[۱۸] Gao and Hitchcock

[۱۹] Cluster

[۲۰] Everitt and Hothorn

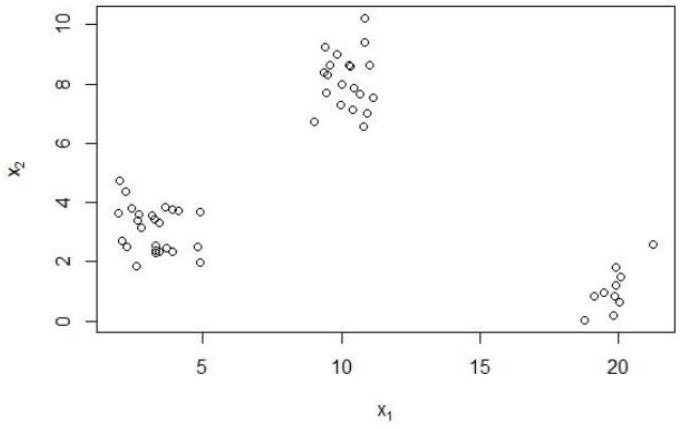
خوشه‌بندی[۲۱]: فرآیند گروه‌بندی (دسته‌بندی) کردن مجموعه‌ای از n نمونه به خوشه‌ها است که تشابه نمونه‌های متعلق به یک خوشه بیشتر از نمونه‌های متعلق به خوشه دیگر است.

تحلیل خوشه‌ای[۲۲]: ابزاری برای اکتشاف ساختار داده‌ها و خوشه‌بندی بدون در نظر گرفتن فرضیات معمول روش‌های آماری است.

الگوریتم‌های خوشه‌بندی[۲۳]: ابزارهایی هستند که در جریان پیدا کردن ساختار داده‌ها در تحلیل خوشه‌ای مورد استفاده قرار می‌گیرند.

عبارت خوشه و تحلیل خوشه‌ای اولین بار در سال ۱۹۳۹ توسط تریان[۲۴] به کار رفته است.

شکل ۱.۱ را در نظر بگیرید:



شکل ۱.۱: داده‌های دومتغیره نشان‌دهنده وجود سه خوشه

بدون نیاز به تفکر و استنتاج، مشاهده می‌کنیم که به طور واضح سه گروه یا خوشه از نقاط وجود دارد. خوشه‌ها به‌وسیله تخمین فاصله نسبی بین نقاط مشخص می‌شوند، در این مثال همگنی نسبی هر خوشه و میزان دوری آنها کار را بسیار ساده می‌کند. زمانی‌که قصد داریم بعضی روش‌های تحلیل خوشه‌ای را برای مجموعه‌ای از داده‌های چندمتغیره به‌کار بریم، آزمون نمودار پراکنش براساس داده‌های اصلی یا حتی مقادیر اولیه مؤلفه‌های اصلی داده‌ها اغلب برای مرحله‌ی ابتدایی بسیار مفید هستند .

[۲۱] Clustering
[۲۲] Clustering analysis
[۲۳] Clustering algorithm
[۲۴] Tryon

خوشه‌بندی، زیرمجموعه‌ای از طبقه‌بندی[۲۵] می‌باشد که به آن یادگیری بدون نظارت[۲۶] نیز گویند و تعداد خوشه‌ها در آن مشخص نیست. به‌عنوان مثال فرض کنید شاخص‌های مختلفی از سلامتی فردی اشخاص سیگاری و غیرسیگاری جمع‌آوری شده است. یک رده‌بندی بدون نظارت، افراد را برمبنای تشابهات بین شاخص‌های سلامتی گروه‌بندی می‌نماید و سپس در سعی است که تعیین کند آیا سیگاری بودن عاملی در سوق دادن افراد به سمت بیماری‌های مختلف بوده است یا خیر. رده‌بندی با نظارت راه‌های تشخیص سیگاری‌ها از غیرسیگاری‌ها را براساس شاخص‌های سلامتی مطالعه می‌کند. در واقع **رده‌بندی بدون نظارت جوهره‌ی تحلیل خوشه‌ای است.** طبقه‌بندی موضوعات متشابه گروه‌ها یک فعالیت مهم است. در زندگی روزمره، این امر قسمتی از فرآیند یادگیری می‌باشد؛ یک کودک تشخیص بین گربه و سگ، میز و صندلی، زن و مرد را یاد می‌گیرد و بدین وسیله رویه طبقه‌بندی ناخودآگاه به‌طور پیوسته بهبود می‌یابد. در سال ۲۰۱۲ مونچه و ویتاکر[۲۷] طبقه‌بندی گسترده‌ای از حیوانات، گیاهان، معادن و بیماری‌ها را تهیه کردند. در علم نجوم دی جاگر[۲۸] و همکاران (۱۹۸۸) ستاره‌ها را براساس دو متغیر فزونی نورشان و دمای سطح آن‌ها، در گروه‌های مختلف طبقه‌بندی کردند. در علوم اجتماعی، خیلی اوقات مردم براساس اولویت‌ها و رفتارشان طبقه‌بندی می‌شوند. در بازاریابی، اغلب بخش‌های بازار مشخص می‌شود (گروه‌هایی از مشتریان با نیازهای متشابه تشکیل می‌شوند). مثال‌های بیشتری در زمینه‌های مختلف جغرافیا (خوشه‌بندی نواحی)، پزشکی (انواع سرطان‌ها)، شیمی (طبقه ترکیبات)، باستان‌شناسی (گروه‌بندی یافته‌های باستان‌شناسی) و غیره می‌توان ارائه داد. .

الگوریتم‌ها و برنامه‌های کامپیوتری زیادی برای تحلیل خوشه‌ای وجود دارد که دلیل تعدد چنین روش‌های مختلف احتمالاً در دو چیز است:

نخست: تحلیل خوشه‌ای از رشد زیادی برخوردار بوده و در بین علوم مختلف کاربرد دارد، که این امر را از هزاران مقاله‌ی چاپ‌شده در مجلات علمی می‌توان فهمید و این باعث شده که پژوهشگران علوم مختلف از رشد آن در علوم دیگر بی‌خبر بمانند. برای همین اطلاعات دقیقی از این‌که توسط چه کسانی توسعه داده شده یا در کدام‌یک از علوم رشد کرده، موجود نیست.

---

[۲۵] Classification

[۲۶] Unsupervised learning

[۲۷] Munsche and Whitaker

[۲۸] De Jager

دوم: تعریف کلی برای خوشه وجود ندارد، و در حقیقت چندین نوع خوشه وجود دارد که عبارتند از: خوشه‌های کروی[۲۹]، خوشه‌های خطی[۳۰]، خوشه‌های محدب[۳۱] و غیره (شکل ۲.۱ را ببینید).

شکل ۲.۱: انواع مختلفی از خوشه‌ها

مجموعه‌ی گسترده‌ای از روش‌ها و همچنین نرم‌افزارهای مختلف موجود، به‌راحتی ذهن کاربر را مشغول این نکته می‌کند که یک دیدگاه مناسب برای حل مسأله انتخاب کند، ولی معیاری که نشان دهد تکنیکی بر تکنیک دیگر برتری دارد موجود نیست و کاربرها از روی مسائل و نوع داده‌ها بهترین روش را انتخاب می‌کنند.

تحلیل خوشه‌ای باید دارای ویژگی‌های زیر باشد:

مقیاس‌پذیری، مناسب بودن برای انواع مختلف از صفات، کشف خوشه‌هایی با شکل‌های مختلف، پیشنیازهای مینیمال برای شناخت زمینه برای تعیین پارامترهای ورودی، توانایی رسیدگی به داده‌های دورافتاده و پرت، حساس نبودن به ترتیب داده‌های ورودی و قابلیت تفسیر و کاربرد (برای آگاهی بیشتر داورزنی، ۱۳۸۹ را ببینید).

همچنین برخی از مشکلات موجود در خوشه‌بندی، به طور خلاصه، عبارتند از:

۱. تکنیک‌های خوشه‌بندی به اندازه کافی پاسخگوی همه نیازها نیست.

۲. بررسی تعداد زیاد داده‌ها با بعد بالا به خاطر پیچیدگی، مشکل‌آفرین است.

[۲۹] Spherical cluster
[۳۰] Linear cluster
[۳۱] Convex cluster

۳. مؤثر بودن روش، به فاصله تعریف‌شده بین واحدها بستگی دارد.

۴. اگر یک اندازه فاصله واضحی وجود نداشته باشد، باید آن را تعریف کرد که تعریف آن، به‌خصوص در فضای چندبعدی کار آسانی نیست.

۵. نتایج الگوریتم‌های خوشه‌بندی می‌تواند، به صورت‌های مختلف قابل تفسیر باشد.

اگر تحلیل خوشه‌ای قادر به تشخیص ساختار داده‌ها نباشد، درعوض ساختاری را به‌وجود می‌آورد که مجموعه داده‌های از هم جداشده با یک روش نسبتاً مناسب، کم و بیش همگن باشند.
در آغاز برای اجرای تحلیل خوشه‌ای، باید داده‌ها را به صورت مناسب نمایش دهیم. در کاربردها، تغییر واحد اندازه‌گیری می‌تواند به ساختار خوشه‌بندی خیلی متفاوتی منجر گردد.

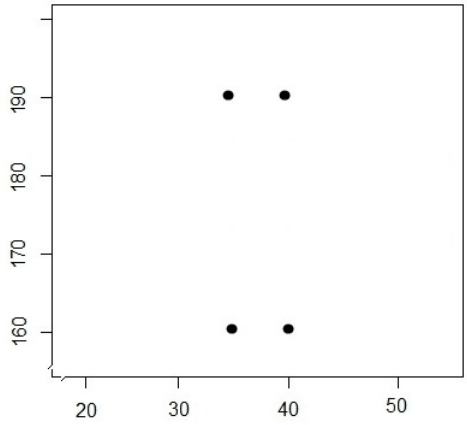
مثال ۱.۲.۱. فرض کنید سن (برحسب سال) و قد (برحسب سانتی‌متر) چهار نفر در جدول ۱.۱ و نمودار دوبعدی آن در نمودار ۳.۱ نمایش داده شده باشند.

جدول ۱.۱: مقادیر قد و سن برحسب سانتیمتر و سال مثال ۱.۲.۱

	صفت	
نفر	سن(yr)	قد(cm)
A	۳۵	۱۹۰
B	۴۰	۱۹۰
C	۳۵	۱۶۰
D	۴۰	۱۶۰

حال اگر قد را برحسب فوت (۱ft=۳۰/۴۸cm) بیان کنیم جدول ۲.۱ و نمودار ۴.۱ را خواهیم داشت.

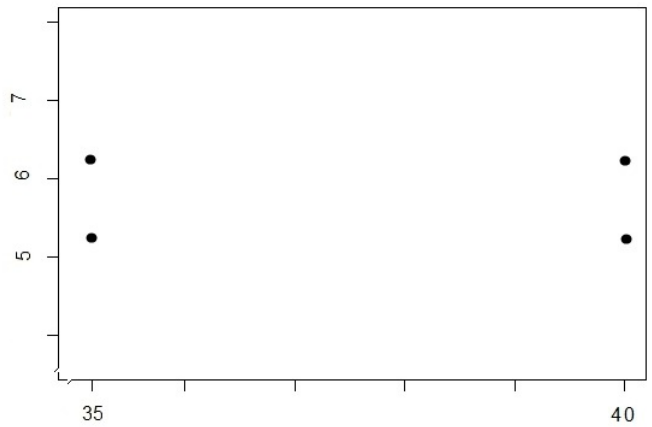
در کاربردها یک راه برای دوری از وابستگی به واحد اندازه‌گیری، می‌تواند استاندارد کردن داده‌ها باشد که با توجه به تعریف آ.۸.۰ صورت می‌گیرد. جدول ۳.۱ مقادیر استانداردشده‌ی جدول ۱.۱ و ۲.۱ می‌باشد و به صورت نمودار ۵.۱ رسم شده است.
نمودار ۵.۱ نشان می‌دهد چهار خوشه داریم، یعنی هر فرد می‌تواند به تنهایی تشکیل یک خوشه دهد یا می‌توانیم همه افراد را عضو یک خوشه در نظر بگیریم. الگوریتم‌های خوشه‌بندی برای انواع مختلف



شکل ۳.۱: نمودار سن و قد جدول ۱.۲.۱

جدول ۲.۱: مقادیر قد و سن برحسب فوت و سال مثال ۱.۲.۱

	صفت	
نفر	سن(yr)	قد(ft)
A	۳۵	۶.۲
B	۴۰	۶.۲
C	۳۵	۵.۲
D	۴۰	۵.۲



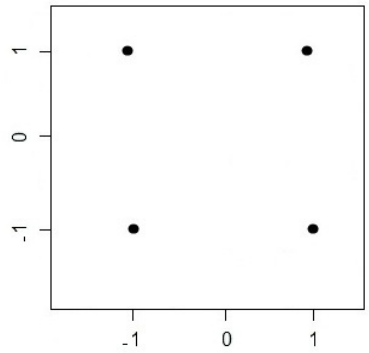
شکل ۴.۱: نمودار سن و قد جدول ۲.۱

داده‌ها سازگارند مگر اینکه عامل‌هایی از قبیل: مقیاس، نرمال کردن و اندازه‌ی مشابهت[۳۲] درست فهمیده

[۳۲] Proximity Measure

جدول ۳.۱: مقادیر استاندارد شده‌ی مثال ۱.۲.۱

	صفت	
نفر	سن(yr)	قد(cm)
A	-۱	۱
B	۱	۱
C	-۱	-۱
D	۱	-۱



شکل ۵.۱: نمودار سن و قد جدول ۳.۱

نشوند و شخص را در تفسیر نتایج دچار اشتباه کنند.

الگوریتم‌های خوشه‌بندی، نمونه‌ها یا داده‌ها را براساس اندازه مشابهت بین دو نمونه گروه‌بندی می‌کنند. مجموعه‌ی نمونه‌ها شامل داده‌های خام هستند که برای تحلیل خوشه‌ای، نیاز به دو صورت استاندارد ماتریس الگو[۳۳] و ماتریس نزدیکی[۳۴] داریم.

۲.۲.۱ ماتریس الگو

نمونه‌ها معمولاً به‌وسیله‌ی بردارهایی (نقاطی) در یک فضای چندبعدی نمایش داده می‌شوند، که هر بعد یک صفت (متغیر) مشخص از نمونه را نشان می‌دهد. اگر فرض کنیم n نمونه داشته باشیم که هر کدام

[۳۳] Pattern matrix

[۳۴] Proximility matrix

دارای d صفت باشند، درحقیقت یک ماتریس نمونه $n \times d$ را می‌توانیم تشکیل دهیم که هر سطر این ماتریس نشان‌دهنده یک نمونه و هر ستون یک صفت مربوط به نمونه می‌باشد. نام دیگر این ماتریس، ماتریس داده می‌باشد. تاکر[۳۵] (۱۹۶۴) عبارت ماتریس دووجهی[۳۶] را به‌کار برده است (چون درایه‌های سطرها و ستون‌ها متفاوت‌اند) و یک نمونه‌ی $n \times d$ دارای ساختار ماتریسی زیر است:

$$\begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{nd} & \dots & x_{nd} \end{pmatrix}$$

۳.۲.۱ ماتریس نزدیکی (مشابهت)

در روش‌های خوشه‌بندی برای ادغام کردن[۳۷] دو خوشه به ماتریسی به نام ماتریس نزدیکی نیازمندیم. این ماتریس که یک ماتریس $n \times n$ می‌باشد (چون سطرها با ستون‌های متناظر یکی هستند) که با استفاده از ماتریس نمونه ساخته شده و مؤلفه‌های سطر iام و ستون jام آن، اندازه‌ی مشابهت[۳۸] یا اندازه‌ی عدم مشابهت[۳۹] می‌باشند (می‌توان از ضریب همبستگی یا فاصله‌ی اقلیدسی به عنوان معیاری برای مشابهت استفاده نمود). به وضوح این ماتریس متقارن است.

مثال ۲.۲.۱. اگر نمونه‌ها تراشه‌های تولیدشده توسط یک دستگاه باشند دو صفت مورد مطالعه می‌تواند عرض و طول تراشه‌ها باشد که جدول ۴.۱ ماتریس نمونه آن را نشان می‌دهد. برای ماتریس الگو اگر اندازه‌ی نزدیکی را فاصله‌ی $p_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ در نظر بگیریم، آنگاه ماتریس نزدیکی به صورت زیر در می‌آید:

[۳۵] Tucker
[۳۶] Two mode
[۳۷] Merge
[۳۸] Similarity measure
[۳۹] Dissimilarity measure

جدول ۴.۱: ماتریس الگوی متناظر چهار نمونه‌ی مثال ۲.۲.۱

	صفت	
تراشه	X	Y
P۱	۱	۲
P۲	۲	۳
P۳	۱	۳
P۴	۲	۴

$$\begin{array}{c} \\ P۱ \\ P۲ \\ P۳ \\ P۴ \end{array} \begin{array}{c} \begin{array}{cccc} P۱ & P۲ & P۳ & P۴ \end{array} \\ \begin{pmatrix} ۰ & \sqrt{۲} & ۱ & \sqrt{۵} \\ \sqrt{۲} & ۰ & ۱ & ۱ \\ ۱ & ۱ & ۰ & \sqrt{۲} \\ \sqrt{۵} & ۱ & \sqrt{۲} & ۰ \end{pmatrix} \end{array}$$

برای تفهیم بهتر ماتریس نزدیکی، اندازه‌ی نزدیکی را بیشتر توضیح می‌دهیم. اگر اندازه‌ی نزدیکی را فاصله‌ی اقلیدسی در نظر بگیریم یک اندازه‌ی عدم مشابهت داریم، بدین صورت که هر چه فاصله‌ی اقلیدسی بین دو نمونه بیشتر باشد عدم مشابهت بین دو نمونه بیشتر شده و هرچه فاصله کمتر باشد عدم مشابهت نیز کمتر می‌شود که در این صورت آن اندازه مشابهت دو نمونه بیشتر می‌گردد.

به طور کلی هر اندازه‌ی نزدیکی بین نمونه‌ی iام و jام باید در سه شرط زیر صدق کند:

۱. (آ) برای عدم مشابهت $\quad dis(i,i) = ۰ \quad \forall i$

(ب) برای مشابهت $\quad sim(i,i) \geq \max_i sim(i,k) \quad \forall (i,k)$

۲. $\quad p(i,k) = p(k,i) \quad \forall (i,k)$

۳. $\quad p(i,k) \geq ۰ \quad \forall (i,k)$

بردار $x_i = (x_{i۱}, x_{i۲}, ..., x_{id})^T$، $i = ۱, ۲, ..., n$، iامین نمونه یا سطر iام ماتریس نمونه می‌باشد.

ماتریس نزدیکی برای داده‌های کمی مناسب بوده و داده‌های کیفی را در برنمی‌گیرد، در حالی‌که ماتریس مشابهت فراگیر است. لذا برای موارد آتی معرفی ماتریس مشابهت لازم به نظر می‌رسد. در این راستا

تنها نیاز داریم اندازه‌ی مشابهت و عدم مشابهت را تعریف کنیم.

اندازه عدم مشابهت

یکی از اندازه‌های مهم عدم مشابهت فاصله‌ی مینکوفسکی[۴۰] است که شکل کلی آن برای نمونه‌ی iام و kام عبارت است از:

$$dis(i,k) = \left(\sum_{j=۱}^{d} |(x_{ij} - x_{kj})|^r \right)^{\frac{۱}{r}} = (||x_i - x_k||^r)^{\frac{۱}{r}}$$

در حالت‌های خاص داریم:

۱. فاصله منهتن[۴۱] $(r = ۱)$

$$dis(i,k) = \sum_{j=۱}^{d} |(xij - xkj)|$$

۲. فاصله اقلیدوسی $(r = ۲)$

$$dis(i,k) = \left[\sum_{ij}^{d} (x_{ij} - xkj)^۲ \right]^{\frac{۱}{۲}} = \left[(x_i - x_k)^T (x_i - x_k) \right]^{\frac{۱}{۲}} = \left(||x_i - x_k||^۲ \right)^{\frac{۱}{۲}}$$

۳. فاصله سوپریمم $(r \to \infty)$

$$dis(i,k) = \sup_{۱ \le j \le d} |(x_{ij} - x_{kj})|$$

اندازه‌ی مشابهت

ضریب همبستگی بین نمونه‌ی iام و jام عبارتست از:

$$|r_{ik}| = \frac{\sum_{j=۱}^{d} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\sqrt{\sum_{j=۱}^{d} (x_{ij} - \bar{x})^۲ \sum_{j=۱}^{d} (x_{kj} - \bar{x}_k)^۲}} \quad , \quad i,k = ۱,...,n$$

که در آن $\bar{x}_l = \frac{۱}{d} \sum_{j=۱}^{d} x_{ij}$ و $l = i,k$ است.

---

[۴۰] Minkowski

[۴۱] Manhatan

۳.۱ روش‌های خوشه‌بندی و الگوریتم‌ها

امروزه روش‌های مختلفی از خوشه‌بندی برای سطوح وسیعی از داده‌ها و همچنین مقاصد مختلف مورد نظر در خوشه‌بندی وجود دارد. خوشه‌بندی زیرمجموعه‌ای از رده‌بندی[۴۲] است و شاخه‌ی مدنظر ما گروه‌بندی بدون نظارت[۴۳] می‌باشد که خود شامل خوشه‌بندی سلسله‌مراتبی، افرازی، مبتنی بر مدل، مبتنی بر شبکه[۴۴] و مبتنی بر چگالی[۴۵] می‌شود. در این بخش به بیان دو روش متداول سلسله‌مراتبی و افرازی می‌پردازیم. روش به‌کار برده شده در این پایان‌نامه خوشه‌بندی K-میانگین و جزء روش‌های افرازی می‌باشد.

۱.۳.۱ خوشه‌بندی سلسله‌مراتبی

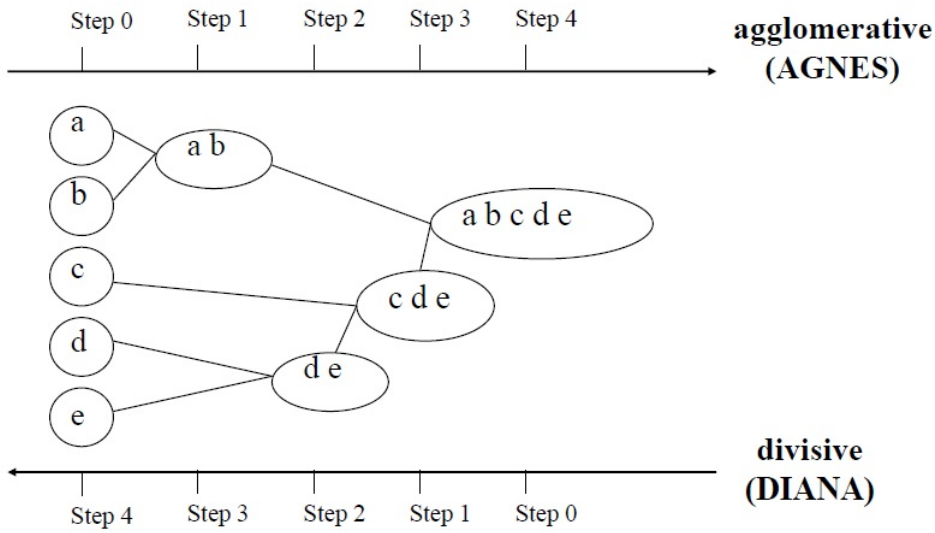
پایه و اساس روش‌های خوشه‌بندی به شیوه‌ی سلسله‌مراتبی، ترکیب کردن خوشه‌ها یا تقسیم آنها می‌باشد. روش خوشه‌بندی سلسله‌مراتبی بر دو نوع تجمعی[۴۶] و تقسیمی[۴۷] می‌باشد. روش‌های سلسله‌مراتبی تجمعی با تمام مشاهدات شروع می‌شود و مشاهداتی که بیشترین تشابه را دارند تشکیل یک گروه می‌دهند. این روند تا زمانی‌که میزان تشابه کاهش یابد ادامه پیدا می‌کند. روش‌های سلسله‌مراتبی تقسیمی برخلاف روش تجمعی است یعنی گروه اولیه از مشاهدات به دو زیرگروه تقسیم می‌شود به‌طوری که مشاهدات هر زیرگروه از مشاهدات گروه دیگر اختلاف زیادی دارد. این گروه‌ها به زیرگروه‌های غیرمشابه تقسیم می‌شوند و این روند تا زمانی‌که هر مشاهده تشکیل یک زیرگروه دهد ادامه می‌یابد. نتایج هر دو روش تجمعی و تقسیمی به صورت نمودار ۶.۱ قابل مشاهده می‌باشد.

در ادامه دنباله افرازهای لانه‌ای را تشریح می‌کنیم.

مجموعه‌ی S از n نمونه به صورت زیر تشکیل شده است:

$$S = \{x_۱, x_۲, \ldots, x_n\}$$

[۴۲] Classificatin
[۴۳] Unsupervised
[۴۴] Grid-based clustering
[۴۵] Density-based clustering
[۴۶] Agglomerative
[۴۷] Divisive



شکل ۶.۱: مقایسه‌ی دو روش تجمعی و تقسیمی

که در آن x_i، iامین نمونه است. یک افراز C از S، S را به زیرمجموعه‌های $\{c_۱, c_۲, \ldots, c_m\}$ تفکیک می‌کند که در شرط زیر صدق می‌کنند:

$$\begin{cases} c_i \cap c_j = \emptyset & \forall i \neq j \quad i,j = ۱,۲,\ldots,m \\ c_۱ \cup c_۲ \cup \ldots \cup c_m = S \end{cases}$$

در نتیجه خوشه‌بندی یک افراز است که مؤلفه‌های این افراز را خوشه می‌نامند. افراز B نسبت به افراز C لانه‌ای است، اگر هر مؤلفه B زیرمجموعه‌ای از یک مؤلفه C باشد، یعنی C از ادغام کردن مؤلفه‌های B تشکیل شده است. برای مثال اگر

$$C = \{(x_۱, x_۳, x_۵, x_۷), (x_۲, x_۴, x_۶, x_۸), (x_۹, x_{۱۰})\}$$

را با سه خوشه و

$$B = \{(x_۱, x_۳), (x_۵, x_۷), (x_۲), (x_۴, x_۶, x_۸), (x_۹, x_{۱۰})\}$$

را با پنج خوشه تعریف کنیم، آنگاه B نسبت به C لانه‌ای است و C و B هر دو یک خوشه‌بندی از مجموعه نمونه‌های

$$\{(x_۱, x_۲, x_۳, x_۴), (x_۵, x_۶, x_۷, x_۸), (x_۹, x_{۱۰})\}$$

می‌باشند.

با توجه به اطلاعات بالا، خوشه‌بندی سلسله‌مراتبی یک دنباله از افرازهایی است که هر افراز نسبت به افراز بعدی لانه‌ای است.

۲.۳.۱ خوشه‌بندی افرازی

نوع دیگر از روش خوشه‌بندی، خوشه‌بندی غیر سلسله‌مراتبی یا روش افرازی می‌باشد. هر دو روش سلسله‌مراتبی و افرازی زمینه‌های کاربرد مخصوص به خود را دارند، روش‌های خوشه‌بندی افرازی فقط نیاز به ماتریس مشابهت دارند، در صورتی‌که روش‌های سلسله‌مراتبی علاوه بر آن به ماتریس الگو نیز نیاز دارند.

روش‌های خوشه‌بندی سلسله‌مراتبی بیشتر در علوم زیست‌شناسی، اجتماعی و رفتاری به دلیل نیاز به ساختن رده‌بندی‌های بسیار پرطرفدار است و تکنیک‌های افرازی در کاربردهای مهندسی که در آن‌ها افرازهای تکی اهمیت دارند مورد استفاده واقع می‌شود. روش‌های افرازی خوشه‌بندی به دلیل آن‌که نیازی به تعیین ماتریس مشابهت ندارد و همچنین داده‌های اصلی در طول اجرای برنامه ذخیره نمی‌شوند برای نمایش مجموعه داده‌های حجیم بسیار مناسب است.

مسأله‌ی خوشه‌بندی افرازی را می‌توان به صورت کلی زیر بیان کرد:

n نمونه در فضای متریک d بعدی مفروض‌اند. افرازی از نمونه‌ها را به صورت k گروه یا خوشه تعیین کنید به طوری‌که نمونه‌های متعلق به یک خوشه بیشتر از نمونه‌های خوشه‌های مختلف دیگر به یکدیگر شبیه باشند. مقدار k تعداد خوشه‌ها، می‌تواند از قبل مشخص باشد یا نباشد. یک معیار خوشه‌بندی مانند توان دوم خطا، باید انتخاب شود. معیارها را می‌توان به صورت فراموضعی یا موضعی طبقه‌بندی کرد. یک معیار فراموضعی در ابتدا k نمونه را انتخاب می‌کند و خوشه‌ها را بدین شکل می‌دهد که هر نمونه‌ای که به یکی از این k نمونه اولیه متشابه‌تر باشد در آن خوشه قرار می‌گیرد؛ ولی معیار موضعی، k خوشه را برحسب تراکم ناحیه‌ها تعیین می‌کند یا k ناحیه‌ای که دارای نزدیک‌ترین همسایگی‌ها هستند. راه حل نظری این مسأله افرازی مستقیم است. یک معیار را انتخاب کرده و آن را به‌ازای کلیه افرازهای ممکن شامل k خوشه ارزیابی کنید و افرازی را که این معیار را بهینه می‌سازد انتخاب نمایید.

اولین مشکلی که خوشه‌بندی افرازی با آن مواجه می‌شود انتخاب یک معیار است که مفاهیم شهودی فرد درباره‌ی خوشه را، به صورت فرمول‌های ریاضی معقول ترجمه کند. معیارها شدیداً به پارامترهای مسأله وابستگی دارند و به دلایل محاسباتی باید ساده، ولی به اندازه‌ی کافی پیچیده باشند تا ساختارهای داده‌های گوناگون را بازتاب دهند.

دومین مشکل این روش ارقام نجومی این افرازها حتی برای تعداد متوسط از داده‌ها می‌باشد، که بدین ترتیب ارزیابی حتی ساده‌ترین معیار نسبت به کلیه افرازها غیر عملی است.

فرض کنید $S(n,k)$ تعداد خوشه‌بندی‌های n نمونه به صورت k خوشه را نشان می‌دهد. از ترتیب نمونه‌ها در هر خوشه و ترتیب خود خوشه‌ها و تعداد خوشه‌های تهی صرف‌نظر شده است. تعداد حالات ممکن برای افراز n نمونه به k خوشه با جواب معادله‌ی تفاضلی جزئی[۴۸] (اعداد استرلینگ[۴۹] از نوع دوم) به‌دست می‌آید:

$$S(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^i\binom{k}{i}(i)^n$$

همانطور که محاسبات نشان می‌دهند، برای افراز ۱۰ نمونه به ۴ خوشه، ۳۴۱۰۵ افراز داریم و این عدد برای افراز ۱۹ نمونه به ۴ خوشه برابر است با ۱۱۲۵۹۶۶۶۰۰۰ که رشد خیلی زیادی داشته است. پس تعیین همه‌ی افرازهای ممکن حتی برای تعداد کم از نمونه‌ها کاری ناممکن است.

کلیات خوشه‌بندی افرازی

فرض کنید n نمونه d بعدی را به K خوشه $\{C_1, C_2, \ldots, C_k\}$ به گونه‌ای افراز کنیم که c_k دارای n_k نمونه و هر نمونه فقط متعلق به یک خوشه است. به طوری‌که $\sum_{k=1}^{K} n_k = n$ و بردار میانگین یا مرکز خوشه c_k به عنوان مرکزیت این خوشه تعریف می‌شود، یعنی

$$c_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{(i)}^{(k)} = \left(C_1^{(k)}, C_2^{(k)}, \ldots, C_d^{(k)}\right)$$

همچنین در آن $x_i^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, \ldots, x_{id}^{(k)})$ نمونه iام متعلق به خوشه c_k است.

$$e_k^2 = \sum_{i=1}^{n_k}(x_i^{(k)} - c_k)^T(x_i^{(k)} - c_k) = \sum_{i=1}^{n_k}||x_i^{(k)} - c_k||^2$$

[۴۸] Partial differnce equation

[۴۹] Stirling numbers

توان دوم خطای خوشه‌ی kام که برابر مجموع توان دوم فاصله‌های اقلیدسی بین هر نمونه در c_k و مرکز خوشه‌اش c_k می‌باشد. این توان‌های دوم خطا تغییرات درون خوشه نیز نامیده می‌شود.
توان دوم فاصله ماهالانوبیس $\{d(i,k) = (x_i, x_k)^T \varphi(x_i, x_k)\}$ که در آن φ ماتریس کواریانس نمونه‌ای آمیخته می‌باشد نیز می‌تواند به عنوان توان دوم خطا به‌کار رود و $E_K^2 = \sum_{k=1}^{K} e_k^2$ مجموع اختلاف بین خوشه‌هاست که باید مینیمم شود.

روش K-میانگین

مک کوئین (۱۹۶۷) عبارت $K-$میانگین را برای توصیف الگوریتمی که هر آیتم را به خوشه‌ای که نزدیکترین مرکز (میانگین) را داشته باشد نسبت می‌دهد. این روش علی رغم سادگی آن یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. این روش الگوریتم‌های مختلفی بیان شده است ولی همه‌ی آنها دارای روالی مشابه هستند که برای تعداد ثابت از خوشه‌ها سعی در به‌دست آوردن موارد زیر را دارند:

- به‌دست آوردن نقاطی به عنوان مراکز خوشه‌ها که در واقع همان میانگین نقاط متعلق به هر خوشه هستند.
- نسبت دادن هر نمونه از داده‌ها به یک خوشه که آن داده کمترین فاصله تا مرکز آن خوشه را دارا باشد.

در نوع ساده‌ای از این روش ابتدا به تعداد خوشه‌های مورد نیاز، نقاطی به صورت تصادفی به عنوان مراکز خوشه انتخاب می‌شوند. سپس بقیه‌ی داده‌ها با توجه به میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شود و بدین ترتیب خوشه‌های اولیه حاصل می‌شود، سپس می‌توان با میانگین‌گیری از داده‌های هر خوشه مراکز جدیدی برای آن خوشه محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که تغییری در خوشه‌ها حاصل نشود.
در ساده‌ترین نوع، این فرآیند از سه مرحله زیر تشکیل می‌شود:

۱. مشخص کردن تعداد k خوشه و سپس تعداد k مشاهده را به‌طور تصادفی به عنوان اولین مرکز خوشه در نظر می‌گیریم. برای تشکیل دادن خوشه‌ها هر مشاهده به خوشه‌ای اختصاص پیدا می‌کند که متناظر با نزدیکترین شیء انتخاب شده تصادفی باشد.

۲. میانگین‌های چندمتغیره (یا مرکز خوشه) خوشه‌ها محاسبه می‌شوند و هر مشاهده به خوشه‌ای که میانگین نزدیک‌تری دارد تخصیص پیدا می‌کند و k خوشه جدید تشکیل می‌شود.

۳. مرحله ۲ تا زمانی‌که میانگین خوشه‌ها از یک مرحله به مرحله بعد تغییر نکنند تکرار می‌شود و سپس الگوریتم متوقف می‌شود.

در روش $K-$میانگین معیار نزدیکی به مرکز خوشه‌ها به وسیله رابطه زیر تعریف می‌شود:

$$d_E^2(x, \bar{x}_c) = (x - \bar{x}_c)'(x - \bar{x}_c) = \sum_k (x_{ik} - \bar{x}_{ck})^2,$$

به طوری‌که $x = (x_1, ..., x_p)'$ یک مشاهده مشخص و $\bar{x}_c$ مرکز خوشه cام می‌باشد.

نسبت دادن نهایی داده‌ها به خوشه‌ها تا اندازه‌ای بستگی به افراز اولیه یا انتخاب اولیه نقاط اصلی (میانگین‌ها) دارد.

برای روشن شدن مطلب مثال ساده‌ی زیر را از حسن‌زاده (۱۳۸۵) مطرح می‌کنیم.

مثال ۱.۳.۱. فرض کنید دو متغیر X_1 و X_2 را برای چهار عنصر A، B، C و D اندازه‌گرفته‌ایم، داده‌ها در جدول ۵.۱ است.

جدول ۵.۱: ماتریس الگو مثال ۱.۳.۱

سیستم	A	B	C	D
X_1	۵	-1	۱	-3
X_2	۳	۱	-2	-2

هدف تقسیم عناصر به $k = 2$ خوشه است به طوری‌که عناصر داخل یک خوشه نزدیک‌تر به آن خوشه نسبت به خوشه دیگر باشد. در پیاده‌سازی روش $K-$میانگین عناصر را به‌طور دلخواه به دو خوشه افراز

می‌کنیم مانند (AB) و (CD) و مولفه‌های $(\bar{x}_۱, \bar{x}_۲)$ را برای میانگین خوشه‌ها محاسبه می‌کنیم. بنابراین در مرحله‌ی ۱ طبق جدول ۶.۱ داریم:

جدول ۶.۱: مؤلفه‌های میانگین در مرحله‌ی اول مثال ۱.۳.۱

خوشه‌ها	$\bar{X}_۱$	$\bar{X}_۲$
AB	$(۵ + (-۱))/۲ = ۲$	$(۳ + ۱)/۲ = ۲$
CD	$(۱ + (-۳))/۲ = -۱$	$((-۲) + (-۲))/۲ = -۲$

مرحله ۲ محاسبه فاصله اقلیدسی برای هر عنصر از مرکز گروه و نسبت دادن به نزدیک‌ترین گروه می‌باشد. اگر هر عنصر از گروه اولیه حرکت داده شود، میانگین خوشه باید مجدداً محاسبه شود.

$$d^۲(A,(AB)) = (۵-۲)^۲ + (۳-۲)^۲ = ۱۰$$

$$d^۲(A,(CD)) = (۵+۱)^۲ + (۳+۲)^۲ = ۶۱$$

A به خوشه (AB) نزدیک‌تر از خوشه (CD) است، پس تغییری در خوشه‌ی عنصر A ایجاد نمی‌شود.

$$d^۲(B,(AB)) = (-۱-۲)^۲ + (۱-۲)^۲ = ۱۰$$

$$d^۲(B,(CD)) = (-۱+۱)^۲ + (۱+۲)^۲ = ۹$$

B به خوشه‌ی (CD) نزدیک‌تر از خوشه‌ی (AB) است، بنابراین به خوشه‌ی نزدیک‌تر منتقل می‌شود. در جدول ۷.۱ خواهیم داشت:

جدول ۷.۱: مؤلفه‌های میانگین در مرحله‌ی دوم مثال ۱.۳.۱

خوشه	$\bar{X}_۱$	$\bar{X}_۲$
A	۵	۳
BCD	-۱	-۱

دوباره هر عنصر را برای نسبت دادن مجدد کنترل می‌کنیم و مقادیر توان دوم فاصله از میانگین هر خوشه را در جدول ۸.۱ داریم.

در نتیجه $k = ۲$ تعداد خوشه‌های نهایی است که (BCD) و A می‌باشند.

جدول ۸.۱: توان دوم فاصله از میانگین هر گروه مثال ۱.۳.۱

خوشه	A	B	C	D
A	۰	۴۰	۴۱	۸۹
BCD	۵۲	۴	۵	۵

مبنای کار روش خوشه‌یابی $K-$میانگین این است که اطلاعات را بر تعداد خوشه‌های از پیش تعیین‌شده، طوری تقسیم می‌کند تا تغییرپذیری اطلاعات درون خوشه حداقل شود. این کار با تغییر مداوم طبقه مربوط به هر بردار از اطلاعات و مراکز طبقات مختلف انجام می‌شود. معمولاً در خوشه‌یابی $K-$میانگین برای اندازه‌گیری میزان نزدیکی داده‌ها به مراکز خوشه‌ها از فاصله اقلیدسی استفاده می‌شود.

مثال ۲.۳.۱. با استفاده از روش $K-$میانگین و ماتریس الگوی ۲.۳.۱ دو خوشه به‌دست آورید.

جدول ۹.۱: ماتریس الگوی مثال ۲.۳.۱

نمونه	صفت اول	صفت دوم
$x_۱$	۰	۲
$x_۲$	۰	۰
$x_۳$	۱/۵	۰
$x_۴$	۵	۰
$x_۵$	۵	۲

ابتدا دو خوشه‌ی $C_۱ = \{x_۱, x_۲, x_۴\}$ و $C_۲ = \{x_۳, x_۵\}$ را به‌طور تصادفی انتخاب می‌کنیم. مرکزیت این دو خوشه عبارتند از:

$$c_۱ = \left(\frac{۰+۰+۵}{۳}, \frac{۲+۰+۰}{۳}\right) = (۱/۶۶, ۰/۶۶)$$
$$c_۲ = \left(\frac{۱/۵+۵}{۲}, \frac{۰+۲}{۲}\right) = (۳/۲۵, ۱/۰۰)$$

اختلاف درون خوشه‌ها

$$e_۱^۲ = [(۰ - ۱/۶۶)^۲ + (۲ - ۰/۶۶)^۲] + [(۰ - ۱/۶۶)^۲ + (۰ - ۰/۶۶)^۲]$$

$$+[(۵ - ۱/۶۶)^۲ + (۰ - ۰/۶۶)^۲] = ۱۹/۳۶$$

$$e_۲^۲ = [(۱/۵ - ۳/۲۵)^۲ + (۰ - ۱)^۲] + [(۵ - ۳/۲۵)^۲ + (۲ - ۱)^۲] = ۸/۱۲$$

برای بررسی تغییرات درون‌خوشه‌ای مجموع توان‌های دوم کل را در هر مرحله محاسبه می‌کنیم.

$$E^۲ = e_۱^۲ + e_۲^۲ = ۱۹/۳۶ + ۸/۱۲ = ۲۷/۴۸$$

حال خوشه‌های جدید طبق جدول ۱۰.۱ عبارتند از $C_۱ = \{x_۱, x_۲, x_۳\}$ و $C_۲ = \{x_۴, x_۵\}$

جدول ۱۰.۱: فاصله‌ی نمونه‌ها از مراکز خوشه‌ها در تکرار اول مثال ۲.۳.۱

نمونه	فاصله از مرکز خوشه‌ی $C_۱$	فاصله از مرکز خوشه‌ی $C_۲$	نزدیکترین خوشه
$x_۱$	۲/۱۴	۳/۴	$C_۱$
$x_۲$	۱/۷۹	۳/۴	$C_۱$
$x_۳$	۰/۸۳	۲/۰۱	$C_۱$
$x_۴$	۳/۴۱	۲/۰۱	$C_۲$
$x_۵$	۳/۶	۲/۰۱	$C_۲$

مرکزیت دو خوشه

$$\begin{cases} c_۱ = (۰/۵, ۰/۶۷) \\ c_۲ = (۵, ۱) \end{cases}$$

اختلاف درون خوشه‌ها

$$\begin{cases} e_۱^۲ = ۴/۱۷ \\ \\ e_۲^۲ = ۲/۰۰ \end{cases} \Rightarrow E^۲ = ۶/۱۷$$

در جدول ۱۱.۱ دیده می‌شود اعضای خوشه‌های $C_۱$ و $C_۲$ تغییر نمی‌کنند و معیار همگرایی به‌دست آمده است.

جدول ۱۱.۱: فاصله‌ی نمونه‌ها از مراکز خوشه‌ها در تکرار دوم مثال ۲.۳.۱

نمونه	فاصله از مرکز خوشه‌ی $C_۱$	فاصله از مرکز خوشه‌ی $C_۲$	نزدیکترین خوشه
$x_۱$	۱/۴۲۱	۵/۱	$C_۱$
$x_۲$	۰/۸۳۶	۵/۱	$C_۱$
$x_۳$	۱/۲۰۴	۳/۶۴	$C_۱$
$x_۴$	۴/۵۵	۱	$C_۲$
$x_۵$	۴/۶۹	۱	$C_۲$

مشکلات روش خوشه‌بندی $K-$میانگین

علی رغم اینکه خاتمه‌پذیری الگوریتم $K-$میانگین تضمین شده است، ولی جواب نهایی آن یکتا نبوده و همواره نتیجه بهینه نمی‌باشد. به طور کلی این روش دارای مشکلات زیر است:

- جواب نهایی به انتخاب خوشه‌های اولیه وابستگی دارد.
- روشی مشخص برای انتخاب مراکز خوشه‌های اولیه وجود ندارد.
- اگر در تکراری از الگوریتم تعداد داده‌های متعلق به خوشه‌ای صفر شد راهی برای تغییر و بهبود ادامه‌ی روش وجود ندارد.
- در این روش فرض بر این است که تعداد خوشه‌ها از ابتدا مشخص باشد اما معمولا در موارد زیادی تعداد خوشه‌ها مشخص نیست.

۴.۱ برآوردگر انقباضی

برآوردگر انقباضی برآوردگری است که از ترکیب برآوردگر خام یا اولیه (MLE، برآوردگر کمترین توان‌های دوم یا بیزی ...) با یک سری اطلاعات بهبود می‌یابد. این بهبود در جهت مقادیر است که از اطلاعات دیگر به‌دست آمده نه در جهت برآوردگر اولیه. در حالت کلی بهبود برآوردگرها را می‌توان با معیار میانگین توان دوم خطا[۵۰] (MSE) اندازه گیری کرد. در این صورت انقباض می‌تواند به سمت صفر یا یک مقدار

[۵۰] Mean squared error

ثابت باشد. اثر این انقباض معمولاً به صورت تبدیل یک برآوردگر نااریب به اریب ولی با MSE کمتر می‌باشد. در عین حال می‌توان برآوردگرهای انقباضی از نوع استاین و نااریب را نیز یافت.

فرض کنید $\theta = (\theta_1, ..., \theta_p)'$ یک پارامتر نامعلوم میانگین با بعد p باشد و $X = (X_1, ..., X_p)'$ بردار مشاهدات p-بعدی به طوری‌که $X \sim N_p(\theta, I_p)$. یافتن بهترین برآوردگر $\hat{\theta}$ بر پایه‌ی مشاهدات X که در آن $\hat{\theta} = \delta(X)$، را در نظر بگیرید. بر اساس توان دوم خطای زیان، سطح کارایی یک برآوردگر $\hat{\theta}$ می‌تواند توسط تابع مخاطره قضاوت شود.

$$R(\theta, \hat{\theta}) = MSE(\hat{\theta}) = E[L(\theta, \hat{\theta})]$$

برای آگاهی بیشتر در این خصوص به نوروزی‌راد (۱۳۹۰) مراجعه کنید.

اگر X یک متغیر تصادفی باشد آن‌گاه شکل کلی برآوردگر انقباضی به صورت $X + g(X)$ است که در آن $g(\cdot)$ یک تابع اندازه‌پذیر می‌باشد. برآوردگر انقباضی بر X برتری دارد اگر

$$\forall \quad \theta \in \Theta \ : \qquad R(\theta, X + g(X)) < R(\theta, X)$$

به‌ازای توابع $g(\cdot)$ مختلف برآوردگرهای متفاوتی به‌دست می‌آید، دو نمونه از آن‌ها عبارتند از:

۱. برآوردگر انقباضی به سمت صفر

$$g(x) = (1 - c)X = X - cX \qquad 0 \leq c \leq 1$$

۲. برآوردگر انقباضی به سمت یک عدد ثابت m

$$g(x) = m + c(X - m) = (1 - c)m + cX \qquad 0 \leq c \leq 1$$

که در آن c ثابت انقباضی است و بسته به نوع توزیع و تابع زیان، مقادیر متفاوتی می‌پذیرد.

در دهه‌های اخیر، رهیافت جیمز-استاین به طور گسترده‌ای در مسائل برآوردهای آماری استفاده شده است. این کار در زمینه برآورد نقطه‌ای میانگین نرمال چندمتغیره آغاز شد.

در سال ۱۹۵۶، چارلز استاین[۵۱] نشان داد که در توزیع نرمال p-متغیره $(p \geq 3)$، برآوردگر طبیعی میانگین نمونه تحت تابع زیان توان دوم، غیرمجاز است. این موضوع سبب ایجاد انگیزه برای بهبود برآوردگرها در مسائل مختلف برآوردیابی شد. سپس جیمز[۵۲] و استاین (۱۹۶۱) برآوردگری ارائه دادند

[۵۱] Stein

[۵۲] James

که بر میانگین نمونه برتری داشت. به عبارت دیگر این برآوردگر دارای مخاطره کمتری نسبت به میانگین نمونه بود. این یافته تحول بزرگی در شاخه‌های مختلف آمار ایجاد کرد. زیرا پس از آن افراد زیادی تلاش کردند برآوردگرهایی از نوع جیمز ـ استاین ارائه دهند که در شرایط مختلف فضای پارامتر دارای مخاطره کمتر باشد.

برآوردگر جیمز ـ استاین برآوردگری غیرخطی است، که بر برآوردگر خطی میانگین نمونه برتری دارد. در نتیجه تمامی خواص خوب برآوردگر میانگین نمونه‌ای، از قبیل ناارییبی و پایایی، که بر اساس نظریه کمترین توان‌های دوم و درستنمایی ماکزیمم بدست می‌آیند، تحت تاثیر برآوردگر جیمز–استاین قرار گرفتند. براین اساس به افتخار پرفسور چارلز استاین، برآوردگری که به‌طور یکنواخت بر برآوردگر استاندارد (برآوردگر کمترین توان‌های دوم LSE و برآوردگر درستنمایی ماکزیمم MLE) برتری دارد را برآوردگر نوع استاین (SE) می‌نامند. فرض کنید $X \in \mathbb{R}^p$ برآوردگر طبیعی پارامتر مکان باشد، ساختار کلی برآوردگرهای نوع استاین مکان به صورت زیر است.

$$\delta_a^{JS}(X) = \Big(1 - \frac{a}{\|X\|^2}\Big)X, \quad 0 < a < 2(p-2)$$

که در آن p بعد فضای پارامتر است. بنابه قضیه گاس–مارکف[۵۳]، برآوردگر LS در رده‌ی برآوردگرهای ناااریب خطی دارای کمترین واریانس است (شائو[۵۴]، ۲۰۰۷) اما ممکن است در ازای اریب شدن برآوردگرها به برآوردگری دست پیدا کنیم که دارای واریانسی (مخاطره) به مراتب کمتر از برآوردگر LS باشد. برآوردگر انقباضی یکی از برآوردگرهایی است که دارای این خاصیت است.

همانطور که مشخص است برآوردگر استاین نیز با فرض $g(X) = -\frac{a}{\|X\|^2}X$، یک برآوردگر انقباضی است. لذا به برآوردگر نوع استاین، برآوردگر انقباضی نوع استاین (SSE) نیز گفته می‌شود.

نتایج استنباط متداول نشان می‌دهد که برآورد درستنمایی ماکسیمم (MLE)، بهترین برآوردگر ناااریب خطی، و برآوردگر کمترین توان دوم همگی برابر میانگین نمونه هستند.

اما استاین به یک پدیده جالب توجه و شگفت انگیز پی برد: اگر $p \leqslant 2$ سپس $\delta_0(X) = X$ مجاز می‌باشد؛ اما اگر $p > 2$، $\delta_0(X) = X$ غیر مجاز است، و $\delta_{JS}(X) = (1 - \frac{p-2}{||X||^2})X$ بر MLE برتری دارد. از آنجا که $\delta_{JS}(X)$ را می‌توان به عنوان یک میانگین وزنی از 0 و X در نظر گرفت، برآوردگر

[۵۳] Gauss-Markow theorem

[۵۴] Shao

جیمز-استاین، برآوردگر انقباضی نامیده می‌شود: $\delta_{JS}(X)$، X را به سمت ۰ منقبض می‌کند.

۱.۴.۱ برآوردگر عمومی جیمز-استاین

برای داده‌های نرمال چندمتغیره ، میانگین نمونه تابع درستنمایی را ماکسیمم می‌کند و واریانس برآوردگر نااریب به‌طور یکنواخت می‌نیمم است $(UMVUE)$. اما جیمز و استاین (۱۹۶۱) نشان دادند که میانگین نمونه غیرمجاز است و برآوردگر آن‌ها، که بعداً برآوردگر جیمز-استاین نامیده شد، بر میانگین نمونه زمانی‌که بعد داده‌ها (p) از ۲ بزرگ‌تر بود غلبه کرد.

برآوردگر عمومی جیمز-استاین وقتی $X \sim N_p(\theta, I_p)$ باشد بدست می‌آید. باک[۵۵] (۱۹۷۵) یک برآوردگر عمومی جیمز-استاین وقتی اعضای X ممکن است همبسته و دارای واریانس متفاوت باشند بدست آورد.

فرض کنید تک مشاهده X دارای توزیع نرمال $p-$متغیره با بردار میانگین θ و ماتریس کواریانس Q باشد، به طوری‌که Q یک ماتریس کواریانس متقارن معین مثبت است. باک نشان داد که یک برآوردگر جیمز-استاین عمومی در این شرایط به صورت زیر است:

$$\delta_{JS}(X) = \left(۱ - \frac{\hat{p} - ۲}{X^t Q^{-۱} X}\right) X$$

که $\hat{p}$ بعد مؤثر[۵۶] ماتریس Q است و برابر

$$\hat{p} = \frac{tr(Q)}{\lambda_{max}(Q)}$$

باک نشان داد که این برآوردگر جیمز-استاین عمومی به $MLE(X)$ تا زمانی‌که $\hat{p} > ۲$ برتری دارد. توجه کنید زمانی‌که $Q = I_p$، داریم $\hat{p} = p$، بعد مؤثر حقیقی می‌شود و برآوردگر عمومی جیمز-استاین تبدیل به برآوردگر اصلی جیمز-استاین می‌شود.

قسمت مثبت برآوردگر جیمز-استاین

این امر که ضریب انقباض ممکن است منفی باشد یک جنبه نامناسب از برآوردگرهای اصلی جیمز-استاین است و می‌توان نشان داد که یک قسمت محدودشده از برآوردگر، برآوردگر برتر است.

[۵۵] Bock

[۵۶] Effective dimension

برای هر اسکالر y ، فرض کنید y^+ قسمت نامنفی y باشد:

$$y^+ = \begin{cases} y, & y \geqslant ۰ \\ ۰, & y < ۰ \end{cases}$$

بنابراین قسمت مثبت برآوردگر جیمز-استاین به صورت زیر است:

$$\delta_{PJS}(X) = \left(۱ - \frac{\hat{p} - ۲}{X^t Q^{-۱} X} \right)^+ X$$

همان‌طور که در لهمن و کاسلا[۵۷] (۱۹۹۸) و ریچارد[۵۸] (۱۹۹۹) نشان داده شده، قسمت مثبت برآوردگر جیمز-استاین به برآوردگر اصلی جیمز-استاین برتری دارد و ما از یک قسمت مثبت برآوردگر جیمز-استاین درون رهیافتمان استفاده می‌کنیم. برآوردگر جیمز-استاین در مهندسی و اقتصاد به‌طور گسترده به‌کار می رود ولی در تحلیل خوشه‌ای توجه کمتری به آن شده است. می‌خواهیم از برآوردگرهای نوع جیمز-استاین به عنوان مرکز خوشه در تحلیل خوشه‌ایی K-میانگین استفاده کنیم به طوری‌که میانگین‌های خوشه‌ها به سمت میانگین کل منقبض شود.

در شرایط معین، ایده‌ی انقباض در تحلیل خوشه‌ای امری رایج است. برای مثال، هیچکاک[۵۹] و همکاران (۲۰۰۷) و هیچکاک و چن[۶۰] (۲۰۰۸) نشان دادند که شیوه‌های انقباض در خوشه‌بندی داده‌های تابعی و داده‌های دودویی به‌کار می‌رود و خوشه‌بندی داده‌های چندمتغیره پیوسته را بهبود می‌بخشد.

فرض کنید بردار مشاهدات $X_{i۱}, X_{i۲}, ..., X_{in_i}$ ، مستقل و هم‌توزیع از توزیع نرمال k-متغیره با بردارهای میانگین μ_i و ماتریس کواریانس Q_i ، که $i = ۱, ..., k$ باشد، یعنی مشاهداتی از k زیرجامعه داریم. میانگین‌های نمونه‌ی k خوشه‌ی تولیدی به‌وسیله‌ی الگوریتم K-میانگین یعنی $\bar{X}_۱, \bar{X}_۲, ..., \bar{X}_k$ و میانگین نمونه کل را $\bar{X}$ در نظر می‌گیریم. مراکز منقبض‌شده جیمز-استاین به‌صورت زیر تعریف می‌شود:

$$\bar{X}_i^{JS} = \bar{X} + \left(۱ - \frac{\hat{p} - ۲}{(\bar{X}_i - \bar{X})^T Q_i^{-۱} (\bar{X}_i - \bar{X})} \right)^+ (\bar{X}_i - \bar{X}) \qquad (۱.۱)$$

سپس از برآوردگرهای انقباضی جیمز-استاین $\bar{X}_i^{JS} (i = ۱, ..., k)$ به عنوان مراکز خوشه در روش K-میانگین استفاده می‌کنیم. توجه کنید که وقتی ماتریس کواریانس زیرجامعه معلوم باشد، مقادیر

[۵۷] Lehmann and Casella
[۵۸] Richard
[۵۹] Hitchcock
[۶۰] Chen

حقیقی Q_i می‌تواند در این فرمول منقبض‌شده استفاده شود. وقتی Q_iها مجهول باشند (حالت رایج در عمل)، از ماتریس کواریانس نمونه درون خوشه‌ای متناظر $\hat{Q}i$ به جای Q_i در فرمول ۱.۱ استفاده می‌شود.

الگوریتم معین مورد استفاده به صورت زیر خلاصه می‌شود.

الگوریتم خوشه‌بندی مبتنی بر جیمز–استاین :

۱. داده‌ها را با استفاده از روش $K-$میانگین با k نقطه آغازین به k خوشه دسته‌بندی کرده و مراکز خوشه‌ها $(\bar{X}_i)$ را به‌دست می‌آوریم.

۲. مراکز خوشه‌های حاصل ($\bar{X}_i$ها) را به سوی میانگین نمونه کل با استفاده از فرمول ۱.۱ منقبض می‌کنیم و مراکز خوشه‌های انقباضی $\bar{X}_i^{JS}$ را به‌دست می‌آوریم.

۳. داده‌ها را با استفاده از روش $K-$میانگین با مراکز خوشه‌های $\bar{X}_i^{JS}$ به k خوشه دسته‌بندی می‌کنیم اگر دو مرکز خوشه انقباضی متمایز نبود، اندکی به آن نوفه‌ی تصادفی اضافه می‌کنیم تا متمایز شوند (به صفحه‌ی ۳۹ مراجعه شود).

تاکید می‌کنیم که ماتریس کواریانس درون خوشه‌ها و Q_i (یا برآورد نمونه‌ای آنها)، تنها در محاسبه مراکز خوشه‌های منقبض‌شده به‌کار می‌رود. فاصله بین هر مشاهده و مرکز هر خوشه‌ای که با قسمت‌بندی توسط الگوریتم $K-$میانگین به‌دست آمده، به صورت فاصله اقلیدسی تعریف می‌شود. مارونا[۶۱] و جاکوکیس[۶۲] (۱۹۷۴) بر روی معیارهای فاصله جایگزین در الگوریتم $K-$میانگین مطالعه کردند و دریافتند که هیچ‌کدام به‌خوبی فاصله اقلیدسی نمی‌باشند. یک روش جایگزین که در این مقاله استفاده نشده است، برآورد مؤلفه کواریانس در $\bar{X}_i^{JS}$ می‌باشد که تمرکز داده‌ها در میان خوشه‌ها صورت می‌گیرد (با تفریق مرکز خوشه مربوط از هر شیء چندمتغیره) و سپس برآورد یک ماتریس کواریانس مشترک با استفاده از تمامی اشیاء.

[۶۱] Maronna

[۶۲] Jacovkis

۵.۱ شاخص رَند

برای ارزیابی دقت خوشه‌بندی به یک روش معین، از شاخص رَند[۶۳] می‌توان استفاده کرد. این شاخص میزان تطابق بین خوشه‌ی حاصل و ساختار خوشه‌بندی واقعی داده‌ها را اندازه‌گیری می‌کند. مسلماً این شاخص در ارزیابی روش خوشه‌بندی با استفاده از شبیه‌سازی مونت کارلو قابل استفاده است، زیرا فرض بر این است که از ساختار خوشه‌بندی واقعی اطلاع داریم. این ملاک عددی در فاصله‌ی $[۰, ۱]$ است. اگر مقدار شاخص رَند یک باشد یعنی تطابق ۱۰۰ درصد می‌باشد و مقدار صفر برای این شاخص به معنی عدم وجود هیچ‌گونه تطابق و بدترین نتیجه‌ی ممکن است ولی در عمل شاخص رَند معمولاً عددی بین ۰ و ۱ می‌گیرد که هر چه به یک نزدیک‌تر باشد خوشه‌بندی صورت گرفته دقیق‌تر است و برعکس. این شاخص که در اصل توسط رَند[۶۴] (۱۹۷۱) تعریف شد و با تان[۶۵] و همکاران (۲۰۰۵) دنبال شد را می‌توان به صورت زیر بیان کرد با فرض این‌که روی N شیء خوشه‌بندی صورت گرفته است، شمارشگرهای زیر تعریف می‌شود:

۱. $N_{۰۰}$ تعداد جفت اشیایی باشد که از یک زیرجامعه متفاوت می‌باشد و با استفاده از روش مورد نظر در خوشه‌ی متفاوتی قرار گرفته است.

۲. $N_{۰۱}$ تعداد جفت اشیایی باشد که از یک زیرجامعه‌ی متفاوت است و با استفاده از روش مورد نظر در خوشه یکسان یا همسان قرار می‌گیرد.

۳. $N_{۱۰}$ تعداد جفت اشیایی که از زیرجامعه‌ی یکسان است و در خوشه‌ی متفاوتی قرار می‌گیرد.

۴. $N_{۱۱}$ تعداد جفت اشیایی است که از زیر جامعه یکسان است و در خوشه یکسان قرار داده می‌شود.

آنگاه

$$Rand = \frac{N_{۰۰} + N_{۱۱}}{N_{۰۰} + N_{۰۱} + N_{۱۰} + N_{۱۱}}$$

[۶۳] Rand
[۶۴] Rand
[۶۵] Tan

این شاخص به عنوان معیار هماهنگی بین ساختار خوشه‌بندی مفروض و نتایج حاصل از روش خوشه‌بندی مورد نظر به کار می‌رود.

برنامه‌نویسی R، برای به‌دست آوردن شاخص رَند براساس تعریف بالا به‌صورت زیر است:

```
N00 <- N01 <- N10 <- N11 <- 0

for (i in 1:N-1) {

for (j in (i+1):N) {

if (U[i]!=U[j]) & (L[i]!=L[j]) N00 <- N00+1

else if (U[i]!=U[j]) & (L[i]==L[j]) N01 <- N01+1

else if (U[i]==U[j]) & (L[i]!=L[j]) N10 <- N10+1

else N11 <- N11+1

}

}

rand <- (N00+N11)/(N00+N01+N1+0N11)
```

که در آن U بردار حاوی شماره‌ی خوشه‌های واقعی و L بردار حاوی شماره‌ی خوشه‌های حاصل از خوشه‌بندی است. به طوری‌که درایه‌ی iام این دو بردار $(i = ۱, \ldots, N)$ هر دو به عضو مشترکی بپردازند. علاوه بر شاخص رند، شاخص رند تعدیل‌شده[۶۶] و شاخص نیمرخ[۶۷] نیز وجود دارند که در این پایان‌نامه به‌کار نرفته‌اند. برای آگاهی بیشتر فرهادی (۱۳۹۳) را ببینید.

[۶۶] Adjusted rand index

[۶۷] Silhouette index

فصل ۲

کاربرد برآوردگر انقباضی جیمز-استاین در خوشه‌بندی K-میانگین

۱.۲ مقدمه

خوشه‌بندی جوامع نرمال یکی از شاخه‌های پرکاربرد خوشه‌بندی است. در دنیای واقعی داده‌های مربوط به این جوامع غالباً چند بعدی هستند، بعد را می‌توان صفت یا ویژگی مشاهدات تفسیر کرد. مشاهداتی از قبیل افراد، اشیاء و ... در صورت مطالعه دارای ویژگی‌های زیادی هستند به‌عنوان مثال ویژگی افراد می‌تواند قد، وزن، فشار خون، ضریب هوش و ... باشد که هر یک دارای توزیع نرمال هستند. همچنین بسیط بودن علوم و دنیای متنوع امروز باعث شده جوامع نرمال بسیاری خود متشکل از چند زیرجامعه‌ی نرمال باشند به‌طوریکه مشاهداتی که متعلق به یک زیرجامعه هستند ویژگی‌های نسبتاً یکسانی دارند. در حالی‌که مشاهدات متعلق به زیرجوامع متفاوت ویژگی‌های نسبتاً متفاوتی دارند. خوشه‌بندی مشاهداتی که دارای چند بعد هستند کار مشکلی است زیرا بایستی تمامی ویژگی‌ها یکسان باشند. با کاربرد رو به رشد خوشه‌بندی جوامع ناهمگن چندبعدی انتظار کارشناسان و متخصصان را برای خوشه‌بندی دقیق‌تر مشاهدات بالا برده است و محققان همواره به‌دنبال روش‌های بهتری بوده‌اند و به خوشه‌بندی $K-$میانگین اکتفا نکرده‌اند.

بعد از این‌که جیمز و استاین نشان دادند در توزیع نرمال p-متغیره $(p \geq ۳)$، برآوردگر طبیعی میانگین نمونه تحت تابع زیان توان دوم، غیرمجاز است عده‌ای به تکاپو افتادند که از این طریق خوشه‌بندی

$K-$میانگین را بهبود بخشند و منقبض کردن مراکز خوشه‌بندی در روش $K-$میانگین یکی از این دست‌آوردها است.

سؤال مهمی که در این قسمت مطرح می‌شود این است که به چه دلیل استفاده از برآوردگر انقباضی ممکن است باعث بهبود خوشه‌بندی شود. اگرچه با یک سری مثال شبیه‌سازی و واقعی سعی می‌کنیم به این سؤال پاسخ دهیم، اما آگاهی رفتار برآوردگرهای انقباضی در مواجهه با برآورد پارامتر میانگین می‌تواند درک بهتری از برتری برای ما ایجاد کند.

در بسیاری از مطالعات نشان داده شده است که برآوردگرهای انقباضی نوع استاین و بارانچیک بهتر از برآوردگر میانگین عمل می‌کنند (برای آگاهی بیشتر لهمن و کسلا، ۱۹۹۸ رابببنید). در ادامه یک سری قضیه به عنوان دلیلی برای ادعای فوق می‌آوریم. لازم به ذکر است که هدف از آوردن این قضایا تنها به منظور نشان دادن برتری برآوردگرهای عنوان شده جهت استفاده در تحلیل خوشه ای است و در هیچ جای دیگر پایان‌نامه به آن‌ها ارجاع داده نمی‌شود.

در اینجا مروری بر مفاهیمی از قبیل لم استاین و روش‌هایی دست‌یابی به برآوردگرهایی بهبودیافته (برآوردگرهای انقباضی) داریم. روش استاین (۱۹۷۲) برپایه‌ی مشخصه‌سازی (تشخیص)[۱] توزیع نرمال استاندارد در ذیل آمده است.

لم ۱.۱.۲. لم استاین (استاین، ۱۹۸۱)

اگر بردار W دارای توزیع نرمال استاندارد باشد، آن‌گاه به ازای تمام توابع پیوسته $g : \mathbb{R} \to \mathbb{R}$ که $E|g'| < \infty$، داریم

$$E[g'(W)] = E[Wg(W)]. \qquad (۱.۲)$$

برهان. فرض کنید g تابعی مطلقاً پیوسته است که $E|g'(z)| < \infty$. اگر W دارای توزیع نرمال

[۱] Characterization

استاندارد باشد، آن‌گاه داریم

$$\begin{aligned} E[g'(W)] &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{\infty} g'(w) e^{-w^۲/۲} dw \\ &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{۰} g'(w) \left(\int_{-\infty}^{w} -x e^{-x^۲/۲} dx \right) dw \\ &+ \frac{۱}{\sqrt{۲\pi}} \int_{۰}^{\infty} g'(w) \left(\int_{w}^{\infty} x e^{-x^۲/۲} dx \right) dw \end{aligned}$$

در این قسمت ترتیب انتگرال‌ها را عوض می‌کنیم.

$$\begin{aligned} E[g'(W)] &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{۰} \left(\int_{x}^{۰} g'(w) dw \right) (-x) e^{-x^۲/۲} dx \\ &+ \frac{۱}{\sqrt{۲\pi}} \int_{۰}^{\infty} \left(\int_{۰}^{x} g'(w) dw \right) x e^{-x^۲/۲} dx \\ &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{۰} \Big(g(۰) - g(x) \Big) (-x) e^{-x^۲/۲} dx \\ &+ \frac{۱}{\sqrt{۲\pi}} \int_{۰}^{\infty} \Big(g(x) - g(۰) \Big) x e^{-x^۲/۲} dx \\ &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{\infty} \Big(g(x) - g(۰) \Big) x e^{-x^۲/۲} dx \\ &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{\infty} x g(x) e^{-x^۲/۲} dx - \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{\infty} x g(۰) e^{-x^۲/۲} dx \\ &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{\infty} x g(x) e^{-x^۲/۲} dx - \frac{g(۰)}{\sqrt{۲\pi}} \int_{-\infty}^{\infty} x e^{-x^۲/۲} dx \\ &= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{\infty} w g(w) e^{-w^۲/۲} dw - ۰ \\ &= E[W g(W)]. \end{aligned} \tag{۲.۲}$$

اثبات کامل است. □

قضیه ۲.۱.۲. فرض کنید $X \sim N_p(\theta, I_p)$ که در آن θ مجهول است. در این‌صورت تحت تابع زیان توان دوم خطا، برآوردگر $\delta_a^{JS}(X)$ بر برآوردگر X برتری دارد اگر $p \geq ۳$، $۰ < a < ۲(p-۲)$ و برآوردگر نسبت به هر برآوردگر $\delta_{p-۲}^{JS} = \left(۱ - \frac{p-۲}{\|X\|^۲}\right) X$ دیگری در این رده به طور یکنواخت دارای کمترین مخاطره است.

برهان. طبق تعریف داریم

$$\begin{aligned}
R(\theta,\delta_a^{JS}) &= E\left[\|\delta_a^{JS}-\theta\|^{۲}\right] = E\left[\|\left(۱-\frac{a}{\|X\|^{۲}}\right)X-\theta\|^{۲}\right] \\
&= E\left[\|\left(X-\theta\right)-\frac{a}{\|X\|^{۲}}X\|^{۲}\right] \\
&= E\left[\|X-\theta\|^{۲}\right]+E\left[\frac{a^{۲}}{\|X\|^{۴}}X'X\right]-۲E\left[\left(X-\theta\right)'\frac{a}{\|X\|^{۲}}X\right] \\
&= E\left[\|X-\theta\|^{۲}\right]+a^{۲}E\left[\frac{X'X}{\|X\|^{۴}}\right]-۲aE\left[\frac{\left(X-\theta\right)'X}{\|X\|^{۲}}\right] \\
&= E\left[\|X-\theta\|^{۲}\right]+a^{۲}E\left[\frac{۱}{\|X\|^{۲}}\right]-۲aE\left[\frac{\left(X-\theta\right)'X}{\|X\|^{۲}}\right] \\
&= p+a^{۲}E\left[\frac{۱}{\|X\|^{۲}}\right]-۲a\sum_{i=۱}^{p}E\left[\frac{\left(X_i-\theta_i\right)X_i}{\|X\|^{۲}}\right]
\end{aligned}$$

با به‌کار بردن لم استاین برای $h_i(X)=\frac{X_i}{\|X\|^{۲}}$، با توجه به این‌که $\frac{\partial\|X\|^{۲}}{\partial X_i}=\frac{\partial\sum X_i^{۲}}{\partial X_i}=۲X_i$ داریم:

$$\frac{\partial h_i(X)}{\partial X_i}=\frac{\|X\|^{۲}-X_i(۲X_i)}{\|X\|^{۴}}$$

لذا می‌توان نتیجه گرفت:

$$\begin{aligned}
R(\theta,\delta_a^{JS}) &= p+a^{۲}E\left[\frac{۱}{\|X\|^{۴}}\right]-۲a\sum_{i=۱}^{p}E\left[\frac{\partial h_i(X)}{\partial X_i}\right] \\
&= p+a^{۲}E\left[\frac{۱}{\|X\|^{۴}}\right]-۲a\sum_{i=۱}^{p}E\left[\frac{\|X\|^{۲}-۲X_i^{۲}}{\|X\|^{۴}}\right] \\
&= p+a^{۲}E\left[\frac{۱}{\|X\|^{۴}}\right]-۲aE\left[\frac{\sum_{i=۱}^{p}\|X\|^{۲}-۲\sum_{i=۱}^{p}X_i^{۲}}{\|X\|^{۴}}\right] \\
&= p+a^{۲}E\left[\frac{۱}{\|X\|^{۴}}\right]-۲aE\left[\frac{p\|X\|^{۲}-۲\|X\|^{۲}}{\|X\|^{۴}}\right] \\
&= p+E\left[\frac{a^{۲}-۲a(p-۲)}{\|X\|^{۲}}\right]=E\left[p+\frac{a^{۲}-۲a(p-۲)}{\|X\|^{۲}}\right]
\end{aligned}$$

از طرفی $R(\theta,X)=p$. بنابراین

$$R(\theta,\delta_a^{JS})-R(\theta,X)\leq ۰ \iff E\left[\frac{a^{۲}-۲a(p-۲)}{\|X\|^{۲}}\right]\leq ۰ \iff$$

$$a^{۲}-۲a(p-۲)\leq ۰ \iff a\left(a-۲\left(p-۲\right)\right)\leq ۰ \iff ۰<a<۲(p-۲)$$

زمانی‌که $a = p - ۲$ است مخاطره‌ی δ_a^{JS} کمترین مقدار است و داریم

$$R(\theta, \delta_{p-۲}^{JS}) = p - E\left[\frac{(p-۲)^۲}{\|X\|^۲}\right]$$

□

قضیه ۳.۱.۲. فرض کنید $X_i \sim N(\theta_i, ۱)$ و X_iها مستقل‌اند، به طوری‌که $۱ \leq i \leq p$ ۰ در این صورت مخاطره‌ی برآوردگر:

$$\delta_c(X) = (۱ - \frac{c(p-۲)}{\| X \|^۲})X$$

تحت تابع زیان $L(\theta, \delta(X)) = \frac{۱}{p}\sum_{i=۱}^{p}(\delta_i - \theta_i)^۲$ به صورت زیر است:

$$R(\theta, \delta_c) = ۱ - \frac{(p-۲)^۲}{p} E_\theta\left[\frac{c(۲-c)}{\|X\|^۲}\right]$$

برهان. طبق تعریف

$$\begin{aligned}
R(\theta, \delta_c) &= E_\theta\left[L\left(\theta, \delta(X)\right)\right] \\
&= E_\theta\left[\frac{۱}{p}\sum_{i=۱}^{p}\left(\delta_i(X) - \theta_i\right)^۲\right] \\
&= E_\theta\left[\frac{۱}{p}\sum_{i=۱}^{p}\left(X_i - \frac{c(p-۲)}{\|X\|^۲}X_i - \theta_i\right)^۲\right] \\
&= \frac{۱}{p}\sum_{i=۱}^{p}\left[E_\theta(X_i - \theta_i)^۲ + E_\theta\left(\frac{c^۲(p-۲)^۲}{\|X\|^۴}X_i^۲\right)\right. \\
&\left. - ۲E_\theta\left((X_i - \theta_i)\left(\frac{c(p-۲)}{\|X\|^۲}X_i\right)\right)\right] \\
&= \frac{۱}{p}\left[p + E_\theta\left(\frac{c^۲(p-۲)^۲}{\|X\|^۲}\right) - ۲\sum_{i=۱}^{p}E_\theta\left((X_i - \theta_i)\left(\frac{c(p-۲)}{\|X\|^۲}X_i\right)\right)\right]
\end{aligned} \tag{۳.۲}$$

حال با فرض $g_i(X) = \frac{c(p-۲)}{\|X\|^۲} X_i$ و استفاده از لم استاین داریم:

$$\begin{aligned} E_\theta\left[(X_i-\theta_i)g_i(X)\right] &= E_\theta\left(\frac{\partial}{\partial X_i} g_i(X)\right) \\ &= E_\theta\left(\frac{\partial}{\partial X_i}\left(\frac{c(p-۲)}{\|X\|^۲}X_i\right)\right) \\ &= E_\theta\left[\frac{c(p-۲)\|X\|^۲ - ۲X_i^۲ c(p-۲)}{\|X\|^۴}\right] \\ &= E_\theta\left[\frac{c(p-۲)}{\|X\|^۲} - \frac{۲X_i^۲ c(p-۲)}{\|X\|^۴}\right] \end{aligned} \tag{۴.۲}$$

سپس با قراردادن رابطه‌ی ۴.۲ در رابطه‌ی ۳.۲ و گرفتن مجموع روی i داریم:

$$\begin{aligned} R\left(\theta, \delta_c(X)\right) &= \frac{۱}{p}\left[p + E_\theta\left(\frac{c^۲(p-۲)^۲}{\|X\|^۲}\right) - ۲E_\theta\left(p\frac{c(p-۲)}{\|X\|^۲} - ۲\frac{c(p-۲)}{\|X\|^۲}\right)\right] \\ &= ۱ + \frac{۱}{p}E_\theta\left[\frac{c^۲(p-۲)^۲}{\|X\|^۲}\right] - \frac{۲}{p}\left[\frac{c(p-۲)^۲}{\|X\|^۲}\right] \\ &= ۱ + \frac{۱}{p}E_\theta\left[\frac{(c^۲-۲c)(p-۲)^۲}{\|X\|^۲}\right] \\ &= ۱ - \frac{(p-۲)^۲}{p}E_\theta\left[\frac{c(۲-c)}{\|X\|^۲}\right] \end{aligned}$$

و اثبات کامل است. □

از قضیه‌ی ۳.۱.۲ نتایج زیر حاصل می‌شود:

نتیجه ۴.۱.۲. اگر $۰ < c < ۲$ و $p \geq ۳$ باشد در این صورت برآوردگر δ_c بر برآوردگر X $(c = ۰)$ وقتی $\delta_c = X$ برتری دارد.

برهان. با توجه به شرط مسأله، $c(۲-c) > ۰$ است لذا مقدار تابع مخاطره با استفاده از قضیه‌ی ۳.۱.۲ از یک کمتر است یعنی به ازای هر θ، داریم $R(\theta, \delta_c) < ۱$. از طرفی برای $c = ۲$ داریم $R(\theta, \delta_c) = R(\theta, X) = ۱$. واضح است که $R(\theta, \delta_c) < R(\theta, X)$ و اثبات کامل است. □

نتیجه ۵.۱.۲. برآوردگر جیمز-استاین δ که مساوی δ_c، با $c = ۱$ است بر همه‌ی برآوردگرهای δ_c وقتی $c \neq ۱$ است برتری دارد.

برهان. با توجه به قضیه‌ی ۳.۱.۲ و رابطه‌ی ۳.۲ داریم

$$R(\theta, \delta_c) = ۱ - \frac{(p-۲)^۲}{p} E_\theta \left[\frac{c(۲-c)}{\|X\|^۲} \right]$$

از طرفی

$$R(\theta, \delta) = ۱ - \frac{(p-۲)^۲}{p} E_\theta \left[\frac{۱}{\|X\|^۲} \right]$$

چون $۰ < c < ۲$ است برای $p \geq ۳$، واضح است که عبارت $c(۲-c)$ مقدار ماکزیمم خود را در $c = ۱$ می‌گیرد. پس $R(\theta, \delta) \leq R(\theta, \delta_c)$ و نتیجه حاصل می‌شود. □

در ادامه به مقایسه‌ی این روش با خوشه‌بندی شناخته شده‌ی $K-$میانگین می‌پردازیم.
این مقایسه از دو جهت انجام می‌شود:

۱. داده‌های ناهمبسته

۲. داده‌های همبسته

با استفاده از شبیه‌سازی مونت کارلو می‌توان به ارزیابی دقت خوشه‌بندی به روش جیمز-استاین پرداخت و در موقعیت‌های یکسان عملکرد خوشه‌بندی جیمز-استاین با سایر انواع خوشه‌بندی را مقایسه کرد.

به طور کلی انجام شبیه‌سازی به این صورت است که از جامعه‌ی مورد نظر که شامل چند زیرجامعه است نمونه‌ای به حجم مشخص گرفته و به خوشه‌بندی اشیاء موجود در آن می‌پردازیم. در نهایت با توجه به این‌که می‌دانیم هر شیء به کدام زیرجامعه تعلق دارد، درصد تطبیق موجود بین واقعیتی که از آن باخبر هستیم و خوشه‌بندی که صورت گرفته است، ملاک قضاوت ما راجع به عملکرد خوشه‌بندی است و البته هر چه تعداد دفعات نمونه‌گیری ما بیشتر باشد ملاک قضاوت ما دقیق‌تر است. از آنجا که جوامع مورد نظر برای خوشه‌بندی جیمز-استاین در این تحقیق دارای زیرجوامع نرمال چندمتغیره هستند به بیان دقیق‌تر و ریاضی‌وار بحث صورت گرفته می‌پردازیم.
مطالب این فصل عمدتاً از مقاله جائو و هیچکاک (۲۰۱۰) برگرفته شده است.

۲.۲ مطالعات شبیه‌سازی

۱.۲.۲ واریانس درون خوشه‌ای متفاوت

داده‌های ناهمبسته

جامعه‌ای متشکل از K توزیع نرمال p-متغیره در نظر بگیرید. با فرض این‌که i معرف زیرجامعه یا توزیع iام است، زیرجامعه‌ی iام دارای بردار میانگین μ_i و ماتریس کواریانس Q_i است، به طوری‌که $i = ۱, \ldots, K$. ماتریس کواریانس زیرجامعه‌ی iام در ساده‌ترین شکل خود $Q_i = \sigma I_p$ می‌باشد، این شکل از ماتریس کواریانس که بیشتر توجه خود را به آن معطوف می‌کنیم به این مفهوم است که پراکندگی درون تمامی زیرجوامع یکسان است.

فرض کنید نمونه‌ای به حجم N از چنین جامعه‌ای گرفته شود با فرض این‌که n_i حجم نمونه‌ی استخراج شده از زیرجامعه‌ی iام است داریم:

$$\sum_{i=۱}^{K} n_i = N$$

در این تحقیق حجم استخراجی از تمامی زیرجوامع را یکسان در نظر می‌گیریم، به این صورت که نمونه‌ای به حجم مساوی از هر یک از زیرجوامع موجود استخراج کرده، روی‌هم می‌ریزیم و سپس به خوشه‌بندی از مجموع این نمونه‌ها، براساس الگوریتم خوشه‌بندی مبتنی بر جیمز-استاین، بصورت زیر می‌پردازیم.

۱. با استفاده از تابع $kmeans$ در R به شکل زیر، الگوریتم K-میانگین دقیقا یک‌بار اجرا شده، مراکز خوشه‌ها و برچسب تمامی داده‌ها در خروجی ظاهر می‌شود.

$$kmeans(X, cl, iter.max = ۱)$$

X ماتریس حاوی مقادیر نمونه‌های حاصل از کل زیر جوامع است. cl در این تابع نشان دهنده‌ی تعداد خوشه‌های موردنیاز و $iter.max$ تعداد تکرارهای الگوریتم K-میانگین را مشخص می‌کند.

۲. مراکز خوشه‌ها با استفاده از رابطه‌ی (۱.۱)، به سمت میانگین کل منقبض می‌شود و مراکز خوشه‌های جدیدی بوجود می‌آید. برای جلوگیری از یکسان شدن مراکز خوشه‌ها و توقف الگوریتم،

در صورتی که حداقل دو مرکز خوشه یکسان باشد، با استفاده از تابع $jitter$ در R به هر مرکز خوشه مقدار اندکی نوفه‌ی تصادفی $y \sim N(۰, ۱۰^{-۵} I_۵)$ می‌افزاییم.

۳. داده‌ها با استفاده از تابع $kmeans$ به‌صورت دستور زیر خوشه‌بندی شده، مراکز خوشه‌های جدید و برچسب تمامی داده‌ها در خروجی ظاهر می‌شود. همچنین با استفاده از برچسب داده‌ها شاخص رند محاسبه می‌شود.

$$kmeans(X, cent, iter.max = ۱)$$

که $cent$ ماتریسی حاوی مراکز خوشه‌ها در مرحله‌ی قبل است.

۴. مراحل ۲ و ۳ آن‌قدر تکرار می شود، تا شاخص رند در دو تکرار پیاپی یکسان گردد، در مواردی همگرایی با این شرط ممکن است بسیار طولانی یا حتی غیرممکن شود، برای جلوگیری از این مشکل، برای تعداد تکرارهای این الگوریتم سقفی تعیین می‌شود (برای مثال ۱۰ تکرار) و در نهایت با الگوریتم K-میانگین با تعداد تکرارهای مساوی مقایسه شود.

در ادامه، الگوی کلی ابتدای بحث برای داده‌های ناهمبسته را با واریانس درون خوشه‌ای متغیر به‌کار می‌گیریم. همان‌طور که گفته شد، ماتریس کواریانس هر زیرجامعه برابر $Q_i = \sigma I_p$، $i = ۱, \ldots, K$ فرض می‌شود. بنابراین تغییر در واریانس درون خوشه‌ای به کمک تغییر در σ صورت می‌گیرد. بنابراین برای σ مقادیر مختلفی در نظر گرفته می‌شود، علاوه بر آن، تعداد زیرجوامع را نیز تغییرمی‌دهیم و اثر آن بر خوشه‌بندی داده‌های ناهمبسته را بررسی می‌کنیم.

مثال ۱.۲.۲. فرض می‌شود جامعه‌ای متشکل از دو زیرجامعه وجود دارد به طوری‌که یکی از زیرجوامع توزیع نرمال ۵-متغیره با بردار میانگین $\delta = (\delta, \ldots, \delta)$ و دیگری توزیع نرمال ۵-متغیره با بردار میانگین $(۰, \ldots, ۰) = ۰$ است و ماتریس کواریانس هر دو زیر جامعه برابر $Q_i = \sigma I_۵$ است $(i = ۱, ۲)$. پارامتر δ پراکندگی بین دو خوشه را مشخص می‌کند در حالی‌که σ، معرف پراکندگی درون هر خوشه است. برای مثال $\delta = ۲$ و σ به‌عنوان پارامتری جهت بررسی اثر تغییر واریانس درون خوشه‌ای بر دقت خوشه‌بندی مقادیر مختلفی چون ۰/۱، ۲، ۴، ... و ۱۰ می‌گیرد. یک حلقه‌ی for حاوی مقادیر σ در نظر می‌گیریم، با فرض این‌که $\sigma = \sigma[j]$ است $(j = ۱, \ldots, ۶)$، ۵۰۰۰ مرتبه نمونه‌ای به حجم $n = ۵۰$ از

هر یک از زیرجوامع مورد نظر می‌گیریم. با استفاده از الگوریتم خوشه‌بندی مبتنی بر جیمز-استاین روی ۱۰۰ عضو نمونه‌ی حاصل به روش‌های مورد نظر (جیمز-استاین و $K-$میانگین سنتی)، خوشه‌بندی انجام می‌دهیم و با استفاده از شاخص رَند میزان دقت خوشه‌بندی را می‌سنجیم، و مقدار آن را ذخیره می‌کنیم.

در انتهای این ۵۰۰۰ مرتبه، میانگین ۵۰۰۰ شاخص رند حاصل از روش خوشه‌بندی مورد نظر را به عنوان میزان دقت خوشه‌بندی مربوطه به ازای واریانس درون خوشه‌ای متناظر محسوب می‌شود. به همین ترتیب این معیار برای تمامی مقادیر σ محاسبه می‌شود. خروجی حاصل از آن‌چه گفته شد در جدول ۱.۲ و نمودار ۱.۲ قابل مشاهده است که حاوی مقایسه بین روش جیمز-استاین معمولی با روش K-میانگین سنتی است.

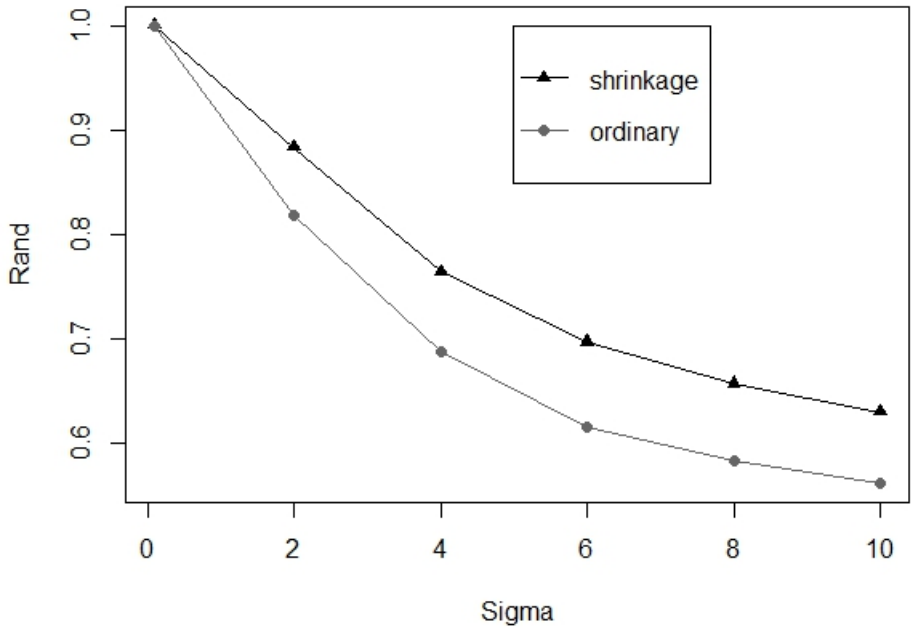
شاخص رند در روش جیمز استاین معمولی، نسبت به شاخص رند در روش K-میانگین سنتی، به ازای

جدول ۱.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از دو روش انقباضی و معمولی برای دو زیرجامعه در مثال ۱.۲.۲

روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$	$\sigma = ۱۰$
انقباضی	۱/۰۰۰۰۰۰۰	۰/۸۸۲۴۸۳۹	۰/۷۶۶۴۰۲۱	۰/۶۹۶۵۴۳۸	۰/۶۵۵۱۳۴۷	۰/۶۳۰۴۴۲۸
معمولی	۰/۹۹۹۷۹۲۳	۰/۸۱۸۸۶۹۷	۰/۶۸۸۹۰۲۴	۰/۶۱۷۰۵۳۹	۰/۵۸۰۸۴۶۷	۰/۵۶۱۷۵۱۲

$\sigma = ۲$ تا $\sigma = ۱۰$ اختلاف معناداری دارد، بنابراین دقت خوشه‌بندی روش جیمز استاین معمولی نسبت به K-میانگین سنتی برتری قابل قبولی دارد.

مثال ۲.۲.۲. این مثال مشابه مثال قبلی است، با این تفاوت که با افزایش خوشه‌های واقعی، مدل پیچیده‌تر شده است. به این‌صورت که به بردارهای میانگین در مثال قبلی سه بردار میانگین جدید اضافه می‌کنیم، در نتیجه بردارهای میانگین ۵ زیرجامعه‌ی تشکیل دهنده‌ی جامعه عبارت است از: $۰ = ۰ \times \mathbf{۱}_۵$، $\mathbf{۲} = ۲ \times \mathbf{۱}_۵$، $-\mathbf{۲} = -۲ \times \mathbf{۱}_۵$، $(-۲, ۰, ۲, ۰, -۲)$ و $(۲, -۲, ۰, -۲, ۲)$، که $\mathbf{۱}_n$ بردار ستونی بطول n است که درایه‌های آن همگی یک هستند. همچنین ساختار ماتریس کواریانس مشابه مثال قبل یعنی $Q_i = \sigma I_۵$ $(i = ۱, \ldots, ۵)$ است، که σ در آن به عنوان متغیر، مقادیر $۰/۲, ۲, ۴, \ldots, ۱۰$ را اختیار



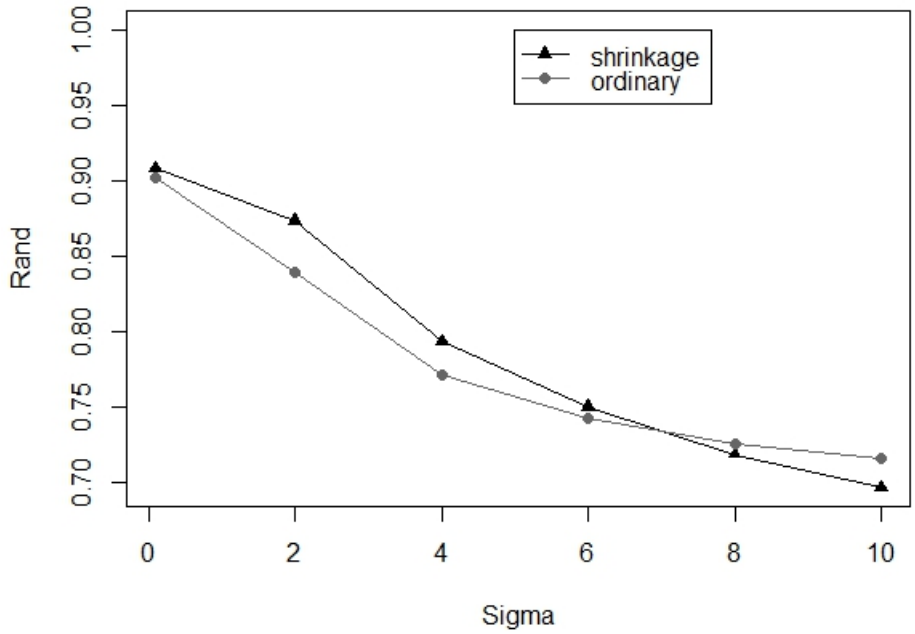
شکل ۱.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از دو زیرجامعه در مثال ۱.۲.۲، بر اساس روش معمولی(دایره) و روش انقباضی(مثلث)

می‌کند. همانند مثال قبل با استخراج ۵۰۰۰ مجموعه داده‌ی ۵۰ تایی (از هر زیرجامعه، ۱۰ نمونه) برای هر $\sigma[j]$ $(j = ۱,\ldots,۶)$ به خوشه‌بندی به دو روش جیمز-استاین و K-میانگین سنتی پرداخته و شاخص رَند (میانگین شاخص رَند برای ۵۰۰۰ بار خوشه‌بندی به ازای $\sigma = \sigma[j]$) را ابزار مقایسه‌ی این دو روش به ازای واریانس درون خوشه‌ای σ قرار می‌دهیم. خروجی حاصل بصورت جدول ۲.۲ و نمودار ۲.۲ است. خروجی به ازای $\sigma \leq ۶$ با آن‌چه انتظار آن می‌رفت مطابقت دارد هر چند که اختلاف

جدول ۲.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از دو روش انقباضی و معمولی برای پنج زیرجامعه در مثال ۲.۲.۲

روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$	$\sigma = ۱۰$
انقباضی	۰/۹۰۷۶۲۲۰	۰/۸۷۳۵۷۸۰	۰/۷۹۲۹۲۴۱	۰/۷۴۸۹۸۸۷	۰/۷۱۹۷۸۷۱	۰/۶۹۵۶۰۱۵
معمولی	۰/۹۰۲۹۳۰۱	۰/۸۳۹۷۳۵۸	۰/۷۷۱۸۲۵۵	۰/۷۴۰۸۲۸۶	۰/۷۲۵۶۳۹۵	۰/۷۱۵۳۲۶۹

شاخص رند به ازای $\sigma = ۶$ ناچیز است، ولی آن‌چه جلب توجه می کند، برتری روش K-میانگین سنتی

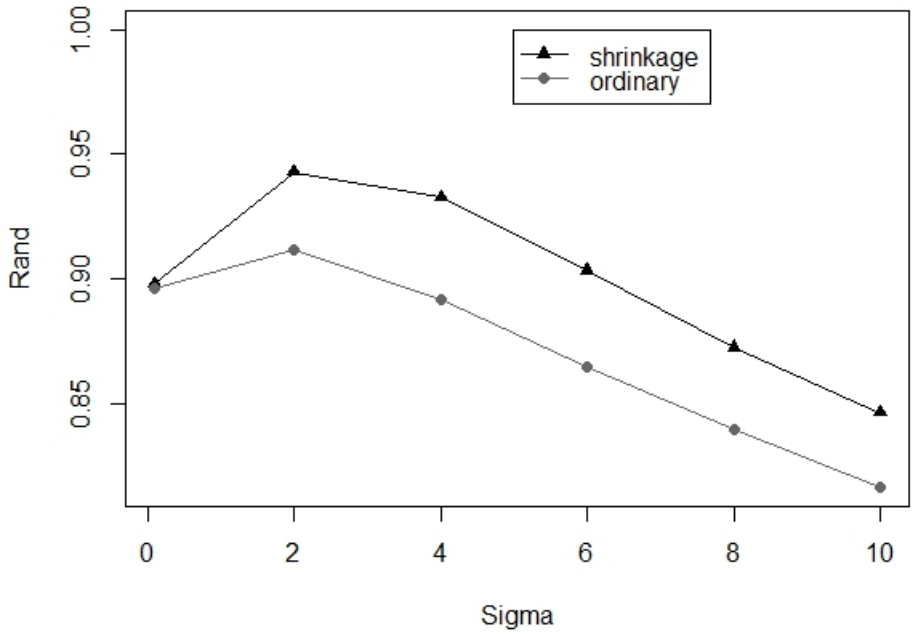


شکل ۲.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از پنج زیرجامعه در مثال ۲.۲.۲، بر اساس روش معمولی(دایره) و روش انقباضی(مثلث)

به ازای $\sigma > 6$ است، که با تصور ما در برتری جیمز-استاین نسبت به K-میانگین به ازای واریانس‌های خوشه‌ای متوسط تا بزرگ مطابقت ندارد.

درست است که جیمز-استاین با افزایش σ نهایتاً برتری خود را از دست می‌دهد ولی در این مثال این افت کمی زودهنگام بود. با ذکر مثال بعد به بررسی بیشتر این مسأله می‌پردازیم.

مثال ۳.۲.۲. این مثال کاملاً شبیه مثال قبل است فقط تغییر کوچکی در بردارهای میانگین ۵ زیرجامعه ایجاد می‌کنیم، به این‌صورت که در بردارهای میانگین مثال قبل به جای مقدار ۲، ۴ و بجای مقدار -2، -4 قرار می‌دهیم. لذا بردارهای میانگین عبارتند از: $۱_۵ \times ۰ = ۰$، $۱_۵ \times ۴ = ۴$، $۱_۵ \times -۴ = -۴$، $(-۴, ۰, ۴, ۰, -۴)$ و $(۴, -۴, ۰, -۴, ۴)$. به بیان دیگر پراکندگی میان خوشه‌ها در این مثال بیشتر از مثال قبل است. این افزایش پراکندگی اثر جالبی بر خروجی‌های به‌دست آمده در جدول ۳.۲ و نمودار ۳.۲ دارد. در نگاه اول می‌بینیم که نتایج هر دو روش افزایش یافته است ولی آن‌چه مدنظر ما است،



شکل ۳.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از پنج زیر جامعه در مثال ۳.۲.۲، بر اساس روش معمولی(دایره) و روش انقباضی(مثلث)

برتری قاطع جیمز-استاین بر $K-$میانگین است.

واقعاً دلیل بهبود نتایج چیست؟ دلیل آن روشن است، اصل و شالوده‌ی روش‌هایی چون جیمز-استاین و حتی $K-$میانگین سنتی مبتنی بر خوشه‌بندی بر اساس کمترین پراکندگی درون خوشه‌ها و بیشترین پراکندگی ممکن بین خوشه‌ها است. بنابراین دلیل این رفتار متفاوت، کمتر شدن واریانس درون خوشه‌ای نسبت به واریانس بین خوشه‌ای است.

جدول ۳.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از دو روش انقباضی و معمولی برای پنج زیرجامعه در مثال ۳.۲.۲

روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$	$\sigma = ۱۰$
انقباضی	۰/۸۹۸۸۸۷۲	۰/۹۴۴۶۵۴۲	۰/۹۳۴۲۸۰۰	۰/۹۰۳۰۲۱۴	۰/۸۷۲۷۴۵۳	۰/۸۴۷۰۸۸۸
معمولی	۰/۸۹۵۶۰۸۳	۰/۹۰۸۴۵۸۸	۰/۸۹۱۷۰۰۴	۰/۸۶۵۰۷۴۹	۰/۸۳۸۷۵۰۹	۰/۸۱۶۳۰۷۶

داده‌های همبسته

علاوه بر آزمون داده‌های که در آن متغیرها ناهمبسته بودند، ما همچنین حالتی که متغیرها همبسته باشند را در نظر گرفتیم، ما داده‌های که دارای ماتریس کواریانس Q با عناصر زیر بودند شبیه‌سازی کردیم.

$$\sigma_{ij} = \begin{cases} \sigma, & i = j \\ r\sigma, & i \neq j \end{cases}$$

نتایج شبیه‌سازی برای $r = ۰/۲۵$ در جدول ۴.۲ داده شده است.

در مقایسه با جدول ۱.۲، یک الگوی مشابهی مشاهده می‌شود، شاخص‌های رند بر اساس روش انقباضی به طور کلی تا حدودی بالاتر از روش سنتی هستند. اما همانطور که در جدول ۱.۲ و ۴.۲ نشان داده شده، بهبود رهیافت انقباضی با داده‌های همبسته کمتر از بهبود با داده‌های ناهمبسته می‌باشد. به عنوان مثال بزرگترین بهسازی برای داده‌های ناهمبسته ۰/۰۷۹۴ است، در حالی‌که بزرگترین بهسازی برای داده‌های همبسته ۰/۰۲۸۵ می‌باشد.

جدول ۴.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از دو روش انقباضی و معمولی برای داده‌های همبسته از دو زیرجامعه

روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$	$\sigma = ۱۰$
انقباضی	۰/۹۹۹۹۰۸۳	۰/۷۵۸۱۵۸۸	۰/۶۵۲۳۰۰۴	۰/۶۰۷۱۷۴۹	۰/۵۸۲۷۵۰۹	۰/۵۶۶۰۰۷۶
معمولی	۰/۹۹۹۵۸۷۲	۰/۷۲۹۶۴۱۲	۰/۶۳۲۴۸۰۰	۰/۵۹۱۲۱۲۴	۰/۵۷۲۰۴۵۳	۰/۵۵۶۴۸۸۸

۲.۲.۲ بعدهای مؤثر مختلف

در اینجا رابطه بین $\hat{p}$ و دقت خوشه‌بندی بررسی می‌شود. در شبیه‌سازی‌های این بخش، مقادیر p و σ به ترتیب برابر ۵ و ۴ در نظر گرفته می‌شوند. داده‌هایی شبیه‌سازی دارای ماتریس کواریانس زیر هستند.

$$Q = \begin{bmatrix} \lambda & ۰ & ۰ & ۰ & ۰ \\ ۰ & \sigma & ۰ & ۰ & ۰ \\ ۰ & ۰ & \sigma & ۰ & ۰ \\ ۰ & ۰ & ۰ & \sigma & ۰ \\ ۰ & ۰ & ۰ & ۰ & \sigma \end{bmatrix}$$

مقادیر جائز برای λ به صورت σ, ۱٫۱۸σ, ۱٫۳۳σ, ۱٫۵۴σ, ۱٫۸۲σ, ۲٫۲۲σ, ۲٫۸۶σ, ۴٫۴σ, ۶٫۶۷σ, ۲۰٫۴σ می‌باشد. در نتیجه $\hat{p}$ برابر ۱٫۲, ۱٫۶, ۲, ۲٫۴, ۲٫۸, ۳٫۲, ۳٫۶, ۴, ۴٫۴, ۵ به‌دست آمد.

نتایج شبیه‌سازی‌های انجام شده در جدول ۵.۲ نشان داده شده است و به طور هندسی در نمودار ۴.۲ رسم شده است. توجه کنید که با افزایش $\hat{p}$ شاخص رند افزایش می‌یابد، زیرا واریانس مؤلفه‌های X

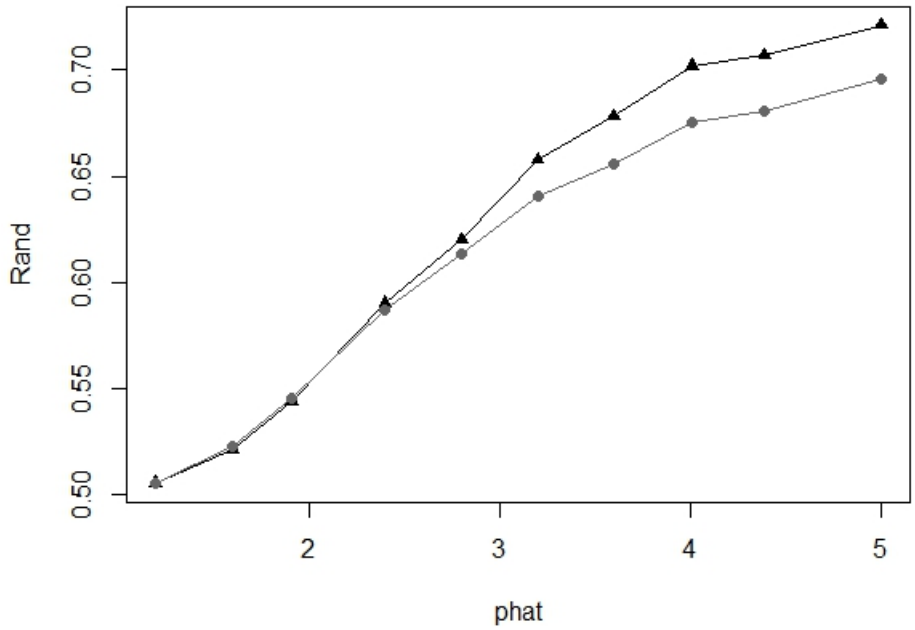
جدول ۵.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از دو روش انقباضی و معمولی برای دو زیرجامعه با تغییر در مقادیر $\hat{p}$

روش	$\hat{p}=$ ۱٫۲	$\hat{p}=$ ۱٫۶	$\hat{p}=$ ۲	$\hat{p}=$ ۲٫۴	$\hat{p}=$ ۲٫۸
انقباضی	۰٫۵۰۵۵۸۷۲	۰٫۵۲۳۳۵۴۲	۰٫۵۵۳۱۸۰۰	۰٫۵۹۵۷۲۱۴	۰٫۶۳۴۰۴۵۳
معمولی	۰٫۵۰۴۸۰۸۳	۰٫۵۲۴۵۵۸۸	۰٫۵۵۳۴۰۰۴	۰٫۵۸۲۸۷۴۹	۰٫۶۰۶۸۵۰۹
روش	$\hat{p}=$ ۳٫۲	$\hat{p}=$ ۳٫۶	$\hat{p}=$ ۴	$\hat{p}=$ ۴٫۴	$\hat{p}=$ ۵
انقباضی	۰٫۶۷۸۵۸۷۲	۰٫۷۱۵۹۵۴۲	۰٫۷۴۱۴۸۰۰	۰٫۷۵۴۲۲۱۴	۰٫۷۶۳۶۴۵۳
معمولی	۰٫۶۲۷۰۰۸۳	۰٫۶۴۳۵۵۸۸	۰٫۶۶۲۲۰۰۴	۰٫۶۷۷۴۷۴۹	۰٫۶۸۴۸۵۰۹

متعادل‌تر می‌باشند و نقاط دور افتاده‌ی کمتری وجود دارد.

در بررسی شاخص‌های رند، زمانی که $\hat{p}$ کمتر یا مساوی ۲ است، ما هیچ تفاوت معنی‌داری بین روش سنتی و انقباضی مشاهده نمی‌کنیم. هنگامی که $\hat{p}$ بزرگتر از ۲ باشد، مقادیر شاخص بر اساس روش انقباضی بالاتر از روش سنتی می‌باشند. علاوه بر این، بهبود از انقباض (که تفاوت بین شاخص‌های رند از این دو روش است) همراه با $\hat{p}$ افزایش می‌یابد، و وقتی که $\hat{p}$ برابر ۵ است به مقدار ماکسیمم خود دست پیدا می‌کند، که این بعد واقعی داده‌های p است.

مشاهده می‌شود نتایج شبیه‌سازی تقریباً سازگار با نتیجه باک در زمینه برآورد هستند؛ $\hat{p}=$ ۲ به نظر می‌رسد نقطه عطف باشد. هنگامی که $\hat{p}\leqslant$ ۲ ، MLE برتر است، بنابراین مراکز خوشه بر پایه MLE نتایج خوشه‌بندی دقیق‌تری را خواهد داد. زمانیکه $\hat{p}>$ ۲ ، برآوردگر انقباضی برتر است و در نتیجه مراکز بر پایه انقباض دقت بیشتری را در نتایج خوشه‌بندی به همراه خواهد داشت.

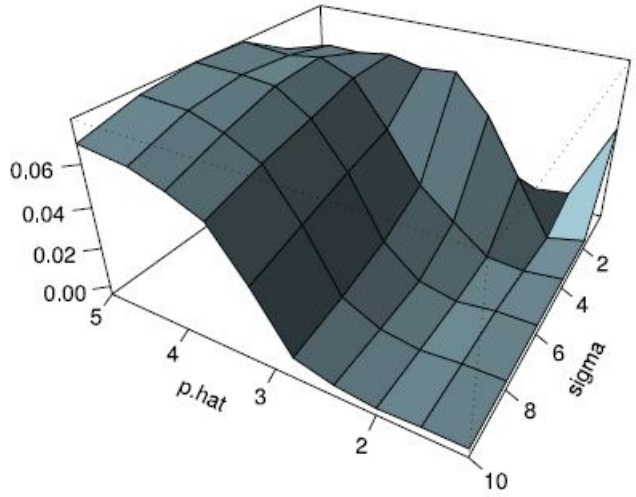


شکل ۴.۲: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از پنج زیر جامعه، تغییرات مختلف $\hat{p}$ بر اساس روش معمولی(دایره) و روش انقباضی(مثلث)

۳.۲.۲ اثر تؤام $\hat{p}$ و σ به طور همزمان

همچنین اثر مشترک $\hat{p}$ و σ روی دقت خوشه‌بندی مورد تحقیق قرارگرفته است، و نتایج شبیه‌سازی انجام شده در نمودار ۵.۲ نمایش داده شده است. نمودار سه بعدی ارتباط بین بهبود (یک بار دیگر به صورت شاخص رند بر اساس روش انقباضی منهای شاخص رند بر اساس روش سنتی تعریف شده) $\hat{p}$ و σ را نشان می‌دهد.

در نمودار سه بعدی، به نظر می‌رسد هرچه $\hat{p}$ افزایش می‌یابد بهبود به طور معنی‌داری افزایش می‌یابد، درحالی‌که بهبود از طریق σ فقط اندکی افزایش می‌کند. بنابراین به نظر می رسد $\hat{p}$ نقش مهمتری نسبت به σ بازی می‌کند. در نمودار سه بعدی، وظیفه منحنی تراز $\hat{p} = ۲$ به عنوان کران واضح است. زمانی‌که $\hat{p} \leqslant ۲$ ، بهبود منفی یا نزدیک به صفر می‌شود. زمانی‌که $\hat{p} > ۲$، هرچه $\hat{p}$ و σ افزایش یابند، بهبود نیز افزایش می‌یابد.



شکل ۵.۲: نمودار بهبود شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ مجموعه داده پنج بعدی از دو زیرجامعه

۳.۲ خوشه‌بندی داده‌های ژن مخمر

در این بخش رهیافت انقباضی را برای داده‌های ژن مخمر به‌کار می‌بریم که توسط آلتر[۲] و همکاران (۲۰۰۰) جمع‌آوری شده است. مجموعه داده‌ها ۷۸ ژن را شامل می‌شود و متغیرها در فاصله‌ی زمانی ۷ دقیقه، ۱۸ بار اندازه‌گیری شده‌اند. جزئیات بیشتر در اسپلمن[۳] و همکاران (۱۹۹۸) شرح داده شده است. همچنین از داده‌ها برای تولید توزیع تقریباً نرمال لگاریتم گرفته شده است.

زیست‌شناسان معتقدند پنج گروه در این ژن‌ها وجود دارد. ژن‌های ۱ تا ۱۳ به گروه $M/G1$، ژن‌های ۱۴ تا ۵۲ به گروه $G1$، ژن‌های ۵۳ تا ۶۰ به گروه S، ژن‌های ۶۱ تا ۶۷ به گروه $S/G2$ و ژن‌های ۶۸ تا ۷۸ به گروه $G2/M$ تعلق دارند. در اینجا حرف S نشانه‌ی کلمه ترکیب[۴] حرف M نشانه‌ی تقسیم سلولی[۵] و حرف G نشانه‌ی شکاف[۶] می‌باشد. در حالی‌که این طبقه‌بندی‌ها به هیچ وجه قطعی نمی‌باشند در این مثال با آن‌ها به عنوان خوشه‌های واقعی رفتار می کنیم.

این داده‌ها را ۷۸ مشاهده‌ی جدا از هم در نظر می‌گیریم و به‌وسیله تابع $kmeans$ در R و استفاده از

[۲] Alter
[۳] Spellman
[۴] Synthesis
[۵] Mitosis
[۶] Gap

برآورگر جیمز-استاین ژن‌ها را به ۵ خوشه تقسیم می‌کنیم. برای داده‌های واقعی این مثال، $\hat{Q}_i$ به عنوان ماتریس کواریانس درون‌خوشه‌ای نمونه به جای Q_i استفاده شده است.

جدول ۶.۲: مقادیر شاخص رند برای داده‌های ژن مخمر با استفاده از روش انقباضی و معمولی

روش	شاخص رند
انقباضی	۰/۷۳۴۳
معمولی	۰/۵۸۱۸

طبق شاخص رَند در جدول ۶.۲، روش K-میانگین معمولی نتوانسته به‌خوبی ساختار خوشه‌بندی تصور شده را اجرا نماید و تنها ۵۸/۱۸ درصد انطباق‌پذیری دارد در صورتی‌که روش انقباضی به قرارگرفتن داده‌ها در یک خوشه‌بندی صحیح نزدیک‌تر شده است و ۷۳/۴۳ درصد داده‌ها به‌وسیله‌ی این روش به‌درستی خوشه‌بندی شده‌اند. از دلایل پایین بودن نسبی شاخص رَند در این مثال می‌توان به عدم تفکیک کامل ژن‌ها از هم اشاره کرد زیرا گروه S (ژن‌های ۵۳ تا ۶۰) تنها خوشه‌ای از پنج خوشه است که جزبه‌جز به‌خوبی تعریف شده و این هشت ژن درون خوشه یکسان قرار گرفته‌اند. گروه $S/G2$ (ژن‌های ۶۱ تا ۶۷) به‌طور ضعیفی تعریف شده و این هفت ژن درون چهار خوشه‌ی مختلف قرار می‌گیرند. سه گروه دیگر به‌طور نسبتاً مناسبی معین شده و درون دو خوشه‌ی مختلف طبقه‌بندی شده‌اند.

در آخر همانطور واضح است رهیافت انقباضی دقت بهتری از رهیافت سنتی به معرض نمایش می‌گذارد.

فصل ۳

برآوردگرهای انقباضی جیمز-استاین تعمیم‌یافته و بارانچیک در خوشه‌بندی K-میانگین

۱.۳ مقدمه

در این فصل، برای مطالعات شبیه‌سازی علاوه بر روش ارائه شده در فصل قبل از دو برآوردگر انقباضی دیگر استفاده می‌کنیم و نتایج حاصل از به‌کار بردن این دو برآوردگر انقباضی را بررسی می‌کنیم.
برآوردگرهای انقباضی گوناگونی وجود دارند. ما در این فصل از دو برآوردگر جیمز-استاین تعمیم‌یافته و بارانچیک[۱] استفاده می‌کنیم که در شبیه‌سازی‌های متعدد نتایج جالب توجهی از خود نشان دادند.
مشابه توضیحات فصل دوم ابتدا در یک سری قضیه که در ادامه می‌آید رفتار و دلیل برتری برآوردگر انقباضی بارانچیک را نسبت به برآوردگر میانگین بررسی می‌کنیم.

قضیه ۱.۱.۳. فرض کنید $X \sim N_p(\theta, I_p)$، $p \geq ۳$، در این صورت برآوردگری به شکل:

$$\delta_i = \left[۱ - \frac{c(\|X\|)(p-۲)}{\|X\|^۲} \right] X_i$$

بارانچیک نامیده می‌شود و تحت تابع زیان

$$L(\theta, \delta) = \frac{۱}{p} \sum_{i=۱}^{p} (\delta_i - \theta_i)^۲$$

[۱] Baranchik

مینیماکس است اگر شرایط زیر برقرار باشد:

(۱) $۰ \leq c(.) \leq ۲$

(۲) تابع $c(.)$، تابعی غیر نزولی باشد.

برهان. ابتدا مخاطره‌ی برآوردگر X را به‌دست می‌آوریم

$$\begin{aligned} R(\theta, X) &= E_\theta \left[\frac{۱}{p} \sum_{i=۱}^{p} (X_i - \theta_i)^۲ \right] \\ &= \frac{۱}{p} \sum_{i=۱}^{p} E_\theta (X_i - \theta_i)^۲ \\ &= \frac{۱}{p} \times p = ۱ \end{aligned}$$

اکنون اگر ثابت کنیم مخاطره‌ی δ_i کوچک‌تر مساوی ۱ است اثبات کامل می‌شود، بدین منظور مخاطره برآوردگر δ را محاسبه می‌کنیم

$$\begin{aligned} R(\theta, \delta) &= E_\theta \left[\frac{۱}{p} \sum_{i=۱}^{p} (\delta_i - \theta_i)^۲ \right] \\ &= E_\theta \left[\frac{۱}{p} \sum_{i=۱}^{p} \left(X_i - \theta_i - \frac{c\left(\|X\|\right)(p-۲)}{\|X\|^۲} X_i \right)^۲ \right] \\ &= \frac{۱}{p} \sum_{i=۱}^{p} \left[E_\theta (X_i - \theta_i)^۲ + E_\theta \left(\frac{c^۲\left(\|X\|\right)(p-۲)^۲}{\|X\|^۲} X_i^۲ \right) \right. \\ &\left. - ۲ E_\theta (X_i - \theta_i) \left(\frac{c\left(\|X\|\right)(p-۲)}{\|X\|^۲} X_i \right) \right] \\ &= \frac{۱}{p} \left[p + E_\theta \left(\frac{c^۲(\|X\|)(p-۲)^۲}{\|X\|^۲} \right) - ۲ \sum_{i=۱}^{p} \left((X_i - \theta_i) g_i(X) \right) \right] \end{aligned} \tag{۱.۳}$$

وقتی $g_i(X) = \frac{c(\|X\|)(p-۲)}{\|X\|^۲}X_i$ می‌باشد. مشتق جزئی این تابع را به‌دست می‌آوریم

$$\begin{aligned}\frac{\partial}{\partial X_i}g_i(X) &= \frac{\partial}{\partial X_i}\left(\frac{c\left(\|X\|\right)(p-۲)}{\|X\|^۲}X_i\right)\\ &= (p-۲)\frac{\left(X_i^۲.\frac{(c'(\|X\|))}{(\|X\|)}+c\left(\|X\|\right)\right)\|X\|^۲ - ۲X_i^۲c\left(\|X\|\right)}{\|X\|^۴}\\ &= (p-۲)\left[\frac{c'\left(\|X\|\right)X_i^۲\|X\|}{\|X\|^۴}+\frac{c\left(\|X\|\right)}{\|X\|^۲}-\frac{۲X_i^۲c\left(\|X\|\right)}{\|X\|^۴}\right]\end{aligned}$$

با مجموع گرفتن روی i داریم

$$\begin{aligned}\sum_{i=۱}^{p}\frac{\partial}{\partial X_i}g_i(X) &= (p-۲)\left[\frac{c'\left(\|X\|\right)}{\|X\|^۲}+\frac{p\left(c\left(\|X\|\right)\right)}{\|X\|^۲}-\frac{۲\left(c\left(\|X\|\right)\right)}{\|X\|^۲}\right]\\ &= (p-۲)\left[\frac{c'\left(\|X\|\right)}{\|X\|^۳}+(p-۲)\frac{c\left(\|X\|\right)}{\|X\|^۲}\right]\end{aligned}$$

در نتیجه بنا به لم استاین

$$-۲\sum_{i=۱}^{p}E_\theta\left((X_i-\theta_i)g_i(X)\right) = -۲(p-۲)E_\theta\left[\frac{c'\left(\|X\|\right)\|X\|}{\|X\|^۲}+(p-۲)\frac{c\left(\|X\|\right)}{\|X\|^۲}\right]$$

بنابراین با جایگذاری در رابطه‌ی ۱.۳ مخاطره‌ی δ به صورت زیر به‌دست می‌آید

$$\begin{aligned}R(\theta,\delta) &= \frac{۱}{p}\Bigg[p+E_\theta\left(\frac{c^۲\left(\|X\|\right)(P-۲)^۲}{\|X\|^۲}\right)-۲(p-۲)E_\theta\left(\frac{c'\left(\|X\|\right)\|X\|}{\|X\|^۲}\right)\\ &\quad -۲(p-۲)^۲E_\theta\left(\frac{c\left(\|X\|\right)}{\|X\|^۲}\right)\Bigg]\\ &= ۱+\frac{۱}{p}E_\theta\left[\frac{(p-۲)^۲c\left(\|X\|\right)}{\|X\|^۲}\left(c\left(\|X\|\right)-\frac{۲\|X\|}{(p-۲)}c'\left(\|X\|\right)-۲\right)\right]\end{aligned}$$

حال چون $۰ \leq c'\left(\|X\|\right)$ و $c\left(\|X\|\right) \leq ۲$ ۰ است عبارت داخل کروشه منفی و مخاطره کمتر از یک می‌شود و در نتیجه این برآوردگر مینیماکس است. □

قضیه ۲.۱.۳. فرض کنید $X \sim N_p(\theta,\Sigma)$ که در آن Σ معلوم است. برآوردگری به شکل:

$$\delta(X) = \left(۱-\frac{c\left(\|X\|^۲\right)}{\|X\|^۲}\right)X$$

برآوردگر بارانچیک تعمیم‌یافته نامیده می‌شود و تحت تابع زیان ، مینیماکس است اگر

(۱) $۰ \leq c\left(\|X\|^۲\right) \leq ۲\left[\frac{tr(\Sigma)}{\lambda_{\max}(\Sigma)}\right] - ۴$

(۲) تابع $c(.)$، غیر نزولی باشد.

که در آن $tr\left(\Sigma\right)$ اثر ماتریس Σ و $\lambda_{\max}\left(\Sigma\right)$ بزرگترین مقدار ویژه‌ی آن است.

برهان. اگر $X \sim N_p(\theta, \Sigma)$ باشد در این صورت

$$\begin{aligned} E_\theta(\theta - X)'(\theta - X) &= E_\theta\left[\sum_{i=۱}^{p}(\theta_i - X_i)^۲\right] \\ &= \sum_{i=۱}^{p} E_\theta(\theta_i - X_i)^۲ \\ &= \sum_{i=۱}^{p} var(X_i) \\ &= \sum_{i=۱}^{p} \sigma_i^۲ \\ &= tr(\Sigma) \end{aligned}$$

از آن‌جایی که برآوردگر X برآوردگر مینیماکس است اگر مخاطره‌ی δ را به‌دست آورده و نشان دهیم تفاضل مخاطره‌ی δ و X مقداری منفی است در این‌صورت برآوردگر δ مینیماکس خواهد بود.

$$\begin{aligned} R(\theta, \delta) &= E_\theta\left[(\theta - \delta(X))'(\theta - \delta(X))\right] \\ &= E_\theta\left[\left(\theta - X - \frac{c\left(\|X\|^۲\right)}{\|X\|^۲}X\right)'\left(\theta - X - \frac{c\left(\|X\|^۲\right)}{\|X\|^۲}X\right)\right] \\ &= E_\theta\left[(\theta - X)'(\theta - X)\right] - ۲E_\theta\left[\frac{c\left(\|X\|^۲\right)}{\|X\|^۲}X'(\theta - X)\right] + E_\theta\left[\frac{c^۲\left(\|X\|^۲\right)}{\|X\|^۲}\right] \\ &= tr(\Sigma) - ۲E_\theta\left[\frac{c\left(\|X\|^۲\right)}{\|X\|^۲}X'(\theta - X)\right] + E_\theta\left[\frac{c^۲\left(\|X\|^۲\right)}{\|X\|^۲}\right] \end{aligned} \tag{۲.۳}$$

عبارت $-۲E_\theta\left[\frac{c\left(\|X\|^۲\right)}{\|X\|^۲}X'(\theta - X)\right]$ را به طریق زیر محاسبه می‌کنیم:

قرار می‌دهیم $\Sigma = PDP'$، که P ماتریس $p \times p$ و متعامد است و D ماتریس قطری که درایه‌های آن مقادیر ویژه‌ی ماتریس Σ می‌باشند.

با فرض $Z = P\Sigma^{-\frac{1}{2}}X$ و $\theta^\star = P\Sigma^{-\frac{1}{2}}\theta$ می‌توان نوشت

$$X = P^{-1}\Sigma^{\frac{1}{2}}Z$$

$$X'X = Z'\Sigma^{\frac{1}{2}}(P^{-1})'P^{-1}\Sigma^{\frac{1}{2}}Z = Z'\Sigma Z = \sum_{i=1}^{p} d_i z_i^2$$

هم چنین

$$\begin{aligned} X'(\theta - X) &= Z'\Sigma^{\frac{1}{2}}(P^{-1})'\left(P^{-1}\Sigma^{\frac{1}{2}}\theta^\star - P^{-1}\Sigma^{\frac{1}{2}}Z\right) \\ &= Z'\Sigma\theta^\star - Z'\Sigma Z \\ &= Z'\Sigma(\theta^\star - Z) \\ &= \sum_{i=1}^{p} d_i z_i(\theta_i^\star - z_i) \end{aligned}$$

پس

$$E_\theta\left[\frac{c\left(\|X\|^2\right)}{\|X\|^2}X'(\theta - X)\right] = \sum_j E_\theta\left[\frac{c\left(\sum_i d_i z_i^2\right)}{\sum_i d_i z_i^2} d_j z_j(\theta_j^\star - z_j)\right]$$

حال با فرض $g_i(Z) = \frac{c\left(\sum_i d_i z_i^2\right)}{\left(\sum_i d_i z_i^2\right) d_j z_j}$ و استفاده از لم استاین داریم

$$\begin{aligned} \sum_j E_\theta\left[g_j(Z)(\theta_j^\star - z_j)\right] &= \sum_j E_\theta\left[\frac{\partial}{\partial z_j} g_j(Z)\right] \\ &= \sum_j E_\theta\Bigg[\frac{\left(\sum_i d_i z_i^2\right)\left(d_j c\left(\sum_i d_i z_i^2\right) + d_j z_j c'\left(\sum_i d_i z_i^2\right) 2 d_j z_j\right)}{\left(\sum_i d_i z_i^2\right)^2} \\ &\quad - \frac{c\left(\sum_i d_i z_i^2\right) d_j z_j 2 d_j z_j}{\left(\sum_i d_i z_i^2\right)^2}\Bigg] \\ &= E_\theta\left[\frac{c\left(\sum_i d_i z_i^2\right)}{\sum_i d_i z_i^2}\sum_j d_j - 2\frac{c\left(\sum_i d_i z_i^2\right)}{\left(\sum_i d_i z_i^2\right)^2}\sum_j d_j^2 z_j^2\right] \\ &\quad + \sum_j E_\theta\left[\frac{c'\left(\sum_i d_i z_i^2\right)}{\sum_i d_i z_i^2} 2 d_j^2 z_j^2\right] \end{aligned}$$

(۳.۳)

اما مقدار $\sum_j d_j^۲ z_j^۲$، برابر است با

$$\begin{aligned}\sum_j d_j^۲ z_j^۲ &= \sum_j z_j d_j^۲ z_j \\ &= Z' D^۲ Z \\ &= Z'(P^{-۱}\Sigma P'^{-۱})^۲ Z \\ &= Z'(P^{-۱}\Sigma P'^{-۱})(P^{-۱}\Sigma P'^{-۱}) Z \\ &= Z'(P^{-۱}\Sigma P'^{-۱} P^{-۱}\Sigma P'^{-۱}) Z \\ &= Z'(P^{-۱}\Sigma^۲ P'^{-۱}) Z \\ &= X'\Sigma^{-\frac{۱}{۲}} P' P^{-۱}\Sigma^۲ P'^{-۱} P \Sigma^{-\frac{۱}{۲}} X \\ &= X'\Sigma X\end{aligned} \tag{۴.۳}$$

و $\sum_j d_j = tr(\Sigma)$ زیرا

$$\begin{aligned}\sum_j d_j = tr(D) &= tr(P^{-۱}\Sigma P'^{-۱}) \\ &= tr(\Sigma P'^{-۱} P^{-۱}) \\ &= tr(\Sigma (PP')^{-۱}) \\ &= tr(\Sigma)\end{aligned} \tag{۵.۳}$$

حال با قراردادن روابط ۴.۳ و ۵.۳ در ۳.۳ داریم

$$\sum_j E_\theta \left[g_j(Z)(\theta_j^\star - z_j) \right]$$

$$= E_\theta \left[\frac{c'\left(\|X\|^۲\right)}{\|X\|^۲} ۲ X'\Sigma X + \frac{c\left(\|X\|^۲\right)}{\|X\|^۲} tr(\Sigma) - ۲ \frac{c\left(\|X\|^۲\right)}{\|X\|^۴} X'\Sigma X \right]$$

بنابراین رابطه‌ی ۲.۳ به صورت زیر به‌دست می‌آید.

$$\begin{aligned}
R(\theta,\delta) &= tr(\Sigma) + E_\theta\left[\frac{c^{۲}\left(\|X\|^{۲}\right)}{\|X\|^{۲}} - ۴\frac{c'\left(\|X\|^{۲}\right)}{\|X\|^{۲}}X'\Sigma X\right.\\
&\quad \left. - ۲\frac{c\left(\|X\|^{۲}\right)}{\|X\|^{۲}}tr(\Sigma) + ۴\frac{c\left(\|X\|^{۲}\right)}{\|X\|^{۴}}X'\Sigma X\right]\\
&= tr(\Sigma) + E_\theta\left[\frac{c\left(\|X\|^{۲}\right)}{\|X\|^{۲}}\left(c\left(\|X\|^{۲}\right) + \frac{۴(X'\Sigma X)}{\|X\|^{۲}} - ۲tr(\Sigma)\right)\right]\\
&\quad - ۴E_\theta\left[\frac{c'\left(\|X\|^{۲}\right)}{\|X\|^{۲}}X'\Sigma X\right]
\end{aligned}$$

چون $c'\left(\|X\|^{۲}\right) \geq ۰$ می‌باشد لذا کافی است که عبارت $c\left(\|X\|^{۲}\right)+۴(X'\Sigma X)\left(\|X\|^{۲}\right)-۲tr(\Sigma)$ منفی باشد تا نتیجه‌ی مورد نظر حاصل شود، پس باید

$$c\left(\|X\|^{۲}\right) < -۴\frac{X'\Sigma X}{\|X\|^{۲}} + ۲tr(\Sigma)$$

اگر داشته باشیم

$$c\left(\|X\|^{۲}\right) < ۲\left[\frac{tr(\Sigma)}{\lambda_{max}(\Sigma)}\right] - ۴$$

در این صورت

$$c\left(\|X\|^{۲}\right)\lambda_{max}(\Sigma) < ۲tr(\Sigma) - ۴\lambda_{max} < -۴\frac{X'\Sigma X}{\|X\|^{۲}} + ۲tr(\Sigma)$$

پس

$$c\left(\|X\|^{۲}\right)\lambda_{max}(\Sigma) + ۴\frac{X'\Sigma X}{\|X\|^{۲}} - ۲tr(\Sigma) < ۰$$

چون $\lambda_{max}(\Sigma) > ۰$، بنابراین

$$c\left(\|X\|^{۲}\right) + ۴\frac{X'\Sigma X}{\|X\|^{۲}} - ۲tr(\Sigma) < ۰$$

لذا تفاضل مخاطره منفی و در نتیجه برآوردگر δ مینیماکس است. □

در ادامه این فصل در بخش ۲.۳، ابتدا عملکرد برآوردگر جیمز-استاین تعمیم‌یافته و در بخش ۳.۳ برآوردگر بارانچیک را در خوشه‌بندی K-میانگین مورد مطالعه قرار می‌دهیم.

۲.۳ خوشه‌بندی با استفاده از برآوردگر جیمز-استاین تعمیم‌یافته

فرض کنید تک مشاهده‌ی $X = (X_۱, X_۲, ..., X_p)$ دارای توزیع نرمال p-متغیره با بردار میانگین $\theta = (\theta_۱, ..., \theta_p)$ و ماتریس کواریانس معین مثبت Q باشد. برآوردگر جیمز-استاین تعمیم‌یافته (استاین ۱۹۵۶) را در نظر بگیرید.

$$\delta_{GJS}(X) = (۱ - \frac{\hat{p} - ۲}{b + X^T Q^{-۱} X})X$$

که

$$\hat{p} = \frac{tr(Q)}{\lambda_{max}(Q)}$$

همانگونه که در فصل قبل بیان شد می‌توان نشان داد برآوردگر محدود شده به جزء مثبت، برتر می‌باشد. برای هر اسکالر Y فرض کنید $Y^+ = \max\{Y, ۰\}$ قسمت نامنفی از Y باشد. بنابراین قسمت مثبت برآوردگر جیمز-استاین تعمیم‌یافته به صورت زیر است :

$$\delta^*_{GJS}(X) = (۱ - \frac{\hat{p} - ۲}{b + X^T Q^{-۱} X})^+ X$$

این برآوردگر نیز، چون برآوردگر جیمزاستاین معمولی، قابلیت به‌کارگیری در تحلیل خوشه‌ای را دارد. به این‌صورت که می‌توان از آن به‌عنوان مرکز خوشه در تحلیل خوشه‌ای به روش $K-$میانگین استفاده کرد، به‌طوری‌که میانگین خوشه‌ها به سمت میانگین کل منقبض می‌شود. حال، فرض کنید $X_i \sim N_p(\theta, Q_i)$، $i = ۱, \cdots, k$. در این‌صورت k زیرجامعه و در نتیجه k خوشه داریم و $\bar{X}_i$، $i = ۱, \ldots, k$، میانگین خوشه‌ی iام تولیدشده توسط الگوریتم K-میانگین است، این میانگین به‌صورت زیر توسط برآوردگر جیمز-استاین تعمیم‌یافته منقبض می‌شود:

$$\bar{X}_i^{GJS} = \bar{X} + \left[۱ - \frac{\hat{p} - ۲}{b + \left(\bar{X}_i - \bar{X}\right)^T Q_i^{-۱} \left(\bar{X}_i - \bar{X}\right)}\right]^+ \left(\bar{X}_i - \bar{X}\right) \qquad (۶.۳)$$

الگوریتم K-میانگین مبتنی بر جیمز-استاین تعمیم‌یافته به جز این‌که در گام ۲، مراکز خوشه‌ها با استفاده از رابطه‌ی فوق منقبض می‌شود، در باقی موارد انطباق کامل با الگوریتم همتای خود مبتنی بر جیمز-استاین معمولی دارد.

در رابطه‌ی ۶.۳، پارامتر b نقش اساسی ایفا می‌کند، زیرا دقت خوشه‌بندی وابسته به مقادیر مختلف b است. در یک خوشه‌بندی تغییر تعداد خوشه‌ها در انجام خوشه‌بندی بر روی مقدار b بهینه، یا bای که

بیشترین دقت را سبب می‌شود، تاثیر می‌گذارد.

همان‌طور که در مثال‌های پیش‌رو خواهیم دید، برآوردگر جیمز-استاین معمولی همواره بر روش K-میانگین سنتی غالب نیست و با افزایش واریانس درون خوشه‌ها، از یک جایی به بعد شروع به افت می‌کند. در این موارد برآوردگر نوع تعمیم‌یافته‌ی آن، به ازای b بهینه، نسبت به K-میانگین سنتی برتری محسوس دارد و در برابر افزایش واریانس استوار است. مثال‌های زیر به فهم مطالب فوق کمک می‌کند.

در این مثال‌ها، به ترتیب تعداد خوشه‌ها افزایش پیدا می‌کند، ولی الگوی یکسانی در همه‌ی آن‌ها وجود دارد. به این‌صورت که نمونه‌ای به حجم n، از جامعه‌ای متشکل از k زیرجامعه‌ی ۵ بعدی، دارای توزیع نرمال تولید می‌شود. میانگین زیرجوامع به‌گونه‌ای طراحی می‌شود که پراکندگی بین خوشه‌ها را نشان می‌دهد. در این مثال‌ها، بررسی خود را بر روی ساده‌ترین ساختار کواریانس جامعه یعنی $Q = \sigma I_۵$ متمرکز می‌کنیم، که σ عددی ثابت است و پراکندگی درون هر خوشه را نشان می‌دهد، البته برای بررسی اثر واریانس درون خوشه، بر دقت خوشه‌بندی به روش‌های مبتنی بر K-میانگین، σ، مقادیر مختلفی می‌گیرد، لذا می‌توان آن را به‌صورت برداری به‌شکل $\sigma = (۰/۱, ۲, ۴, \ldots, ۲۰)$ درنظر گرفت. واضح است که کواریانس میان صفات صفر فرض می‌شود. روش کار در این برآوردگر کاملا شبیه به برآوردگر جیمز-استاین است، با این تفاوت که در برآوردگر جیمز-استاین معمولی، به‌عنوان حالت خاصی از برآوردگر جیمز-استاین تعمیم‌یافته $b = ۰$ است. مقادیر متنابهی نمونه به حجم n، از جامعه‌ی نام‌برده تولید شده، به ازای مقادیر مختلف b، میانگین شاخص رند این نمونه‌ها محاسبه می‌شود. در نهایت بین مقادیر مختلف b، برآوردگری که بیشترین دقت را در میان بقیه احراز می‌کند، به همراه برآوردگرهای جیمز-استاین معمولی و K-میانگین سنتی، در خروجی نمایش داده می‌شود.

به بیان دقیق‌تر، همچون مقایسه‌ی بین برآوردگر مبتنی بر جیمز-استاین و K-میانگین معمولی در مثال‌های فصل قبل، به ازای b مفروض، σ به‌صورت مجموعه مقادیر $\{۰/۱, ۲, ۴, \ldots, ۲۰\}$ در نظر گرفته می‌شود، که با استفاده از شاخص رند، دقت خوشه‌بندی به دو روش جیمز-استاین با پارامتر b و K-میانگین سنتی به ازای مقادیر مختلف σ بررسی می‌شود. همان‌طور که می‌دانید، این بررسی و مقایسه به ازای هر σ_i، $i = ۱, \ldots, ۱۱$، به کمک شاخص رند انجام می‌گیرد، که برای بالا بردن دقت این شاخص به ازای b مفروض و هر σ_i، $i = ۱, \ldots, ۱۱$، ۱۰۰ مرتبه نمونه‌ای به حجم n به تعداد

مساوی از k زیرجامعه‌ی ۵ بعدی دارای توزیع نرمال با میانگین‌های مفروض و با کواریانس جامعه برابر $Q_i = \sigma_i I_۵$، $i = ۱, \ldots, ۱۱$، گرفته، روی آن به دو روش جیمز-استاین با پارامتر b و K-میانگین سنتی خوشه‌بندی کرده، از میانگین شاخص رندهای حاصل از این ۲۰۰ مرتبه، به عنوان جایگزین مناسب‌تر استفاده می‌شود.

برای اعمال سایر مقادیر b، با استفاده از تعریف برداری چون $b = (۰, ۱, ۱۰, ۱۰۰, ۱۰۰۰, \hat{p} - ۲)$ و یک حلقه‌ی for در R برای قراردادن هریک از از این مقادیر به‌جای b مفروض، مطالب فوق را از مقداری ثابت برای b، به برداری حاوی مقادیر b تعمیم داد، و به مقایسه‌ی عملکرد برآوردگر متناظر با هر b_j، $j = ۱, \ldots, ۶$ پرداخت.

مثال ۱.۲.۳. نمونه‌ای به حجم $n = ۵۰$، از جامعه‌ای متشکل از دو زیرجامعه‌ی ۵ بعدی، دارای توزیع نرمال تولید می‌شود. میانگین زیرجوامع $\delta = (\delta, \delta, \delta, \delta, \delta)^T$ و $۰ = (۰, ۰, ۰, ۰, ۰)^T$ فرض می‌شود، که δ ثابت است. همچنین طبق قراردادی که بیان شد، کواریانس جامعه $Q_i = \sigma_i I_۵$، $i = ۱, \ldots, ۱۱$، درنظرگرفته می‌شود، که σ_i مولفه‌ی i ام بردار σای است، که پیش از طرح این مثال تعریف شد.

استفاده از میانگین شاخص‌های رند حاصل از ۱۰۰ بار نمونه‌گیری به حجم $n = ۵۰$ به تعداد یکسان از هر دو زیرجامعه‌ی موجود به ازای هر σ_i، $i = ۱, \ldots, ۱۱$ و به ازای هر b_j، $j = ۱, \ldots, ۶$، همان‌گونه که در جدول ۱.۳ می‌آید، مقایسه‌ی برآوردگرهای متناظر با هر b_j، $j = ۱, \ldots, ۶$، برحسب مقادیر مختلف σ را ممکن می‌سازد.

مقادیر شاخص رند به ازای تمامی مقادیر b، در تمامی مقادیر σ نسبت به برآوردگر K-میانگین سنتی برتری دارد. در میان مقادیر b، مقدار $b = ۰$، که متناظر با برآوردگر جیمز-استاین معمولی است، از سایر مقادیر بهتر است. نمودار ۱.۳ این برآوردگر را در کنار برآوردگر متناظر با $b = ۱$ و برآوردگر K-میانگین سنتی به‌طور شهودی مقایسه می‌شود.

مثال ۲.۲.۳. این مثال با ایجاد یک سری تغییرات جزئی در مثال قبلی بوجود آمده است، و باقی مطالب همچنان به قوت خود باقی است. جامعه‌ی مورد بررسی متشکل از ۵ زیرجامعه است و در نتیجه تشکیل ۵ خوشه منطقی به‌نظر می‌رسد. هر زیرجامعه دارای توزیع نرمال ۵ متغیره با بردار میانگین مخصوص خود

جدول ۱.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای دو زیرجامعه در مثال ۱.۲.۳

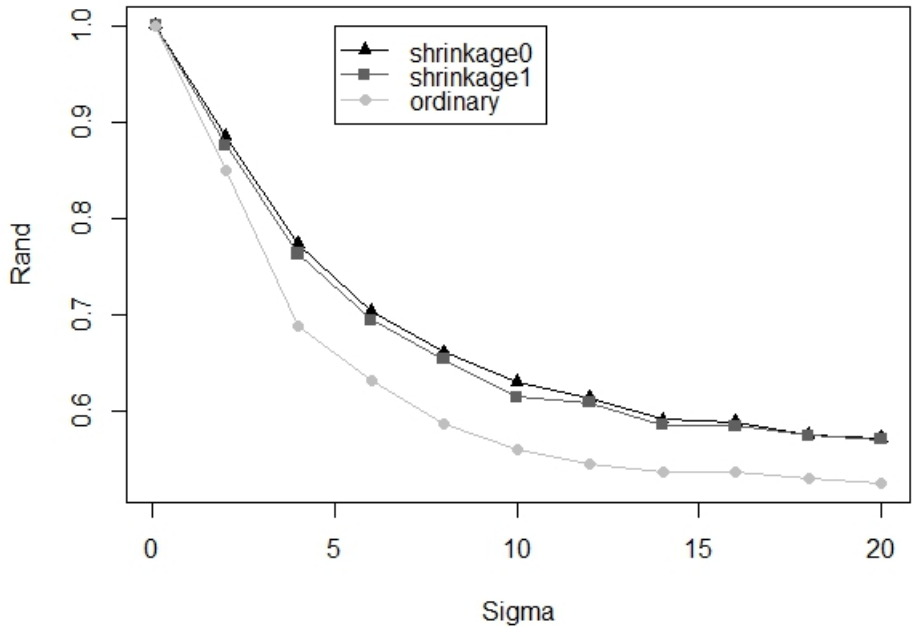
روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$
تعمیم یافته	۱/۰۰۰۰۰۰۰	۰/۸۷۷۱۵۹۲	۰/۷۶۰۷۱۰۲	۰/۷۰۰۳۹۱۸	۰/۶۴۵۶۹۸۰
انقباضی	۱/۰۰۰۰۰۰۰	۰/۸۸۶۹۴۶۹	۰/۷۶۷۹۹۱۸	۰/۷۰۰۸۱۶۳	۰/۶۵۱۷۸۷۸
معمولی	۱/۰۰۰۰۰۰۰	۰/۸۴۳۱۱۸۴	۰/۶۹۸۵۸۷۸	۰/۶۲۰۵۷۱۴	۰/۵۸۰۱۷۱۴
$\sigma = ۱۰$	$\sigma = ۱۲$	$\sigma = ۱۴$	$\sigma = ۱۶$	$\sigma = ۱۸$	$\sigma = ۲۰$
۰/۶۲۸۸۸۱۶	۰/۶۲۸۸۸۱۶	۰/۵۹۰۲۶۱۲	۰/۵۷۶۵۸۷۸	۰/۵۷۴۶۶۹۴	۰/۵۵۷۰۸۵۷
۰/۶۳۶۳۹۱۸	۰/۶۱۵۵۱۰۲	۰/۵۹۴۹۵۵۱	۰/۵۸۲۷۸۳۷	۰/۵۷۹۸۰۴۱	۵۶۰۱۴۶۹
۰/۵۵۹۱۹۱۸	۰/۵۴۲۳۷۵۵	۰/۵۴۰۶۱۲۲	۰/۵۲۷۳۴۶۹	۰/۵۲۵۰۲۰۴	۰/۵۱۸۹۲۲۴

و ماتریس کواریانس واریانس برابر $Q = \sigma I_۵$ است. این بردار میانگین‌ها عبارتند از: $۰ = ۰ \times ۱_۵$، $۲ = ۲ \times ۱_۵$، $-۲ = -۲ \times ۱_۵$، $۴ = ۴ \times ۱_۵$ و $-۴ = -۴ \times ۱_۵$. بردار ۱_n برداری بطول n که تمامی درایه‌های آن ۱ است.

سپس خوشه‌بندی به دو روش جیمزاستاین تعمیم‌یافته و $K-$میانگین معمولی، براساس مقادیر مختلف σ، یعنی $\{۰/۱, ۲, ۴, \ldots, ۱۰\}$ انجام می‌گیرد. که روش جیمزاستاین تعمیم‌یافته، شامل مقادیر مختلف بردار $b = (۰, ۱, ۱۰, ۱۰۰, ۱۰۰۰, \hat{p}-۲)$ است. خروجی حاصل از خوشه‌بندی براساس مقادیر مختلف σ و مقادیر مختلف b بصورت جدول ۲.۳ و نمودار ۲.۳ است. در خروجی‌های حاصل، بین مقادیر b،

جدول ۲.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای پنج زیرجامعه در مثال ۲.۲.۳

روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$
تعمیم یافته	۰/۹۲۴۸۰۰۰	۰/۸۶۶۰۹۸۰	۰/۸۲۰۶۱۲۲	۰/۷۹۷۷۹۵۹	۰/۷۷۹۶۴۰۸
انقباضی	۰/۹۲۴۸۱۶۳	۰/۸۶۶۰۴۹۰	۰/۸۱۹۷۴۶۹	۰/۷۹۹۷۸۷۸	۰/۷۸۰۴۳۲۷
معمولی	۰/۹۲۹۸۲۰۴	۰/۸۴۸۶۷۷۶	۰/۸۰۷۱۷۵۵	۰/۷۸۰۳۱۸۴	۰/۷۶۷۰۳۶۷
$\sigma = ۱۰$	$\sigma = ۱۲$	$\sigma = ۱۴$	$\sigma = ۱۶$	$\sigma = ۱۸$	$\sigma = ۲۰$
۰/۷۷۰۳۲۶۵	۰/۷۶۱۲۷۳۵	۰/۷۵۶۳۶۷۳	۰/۷۵۰۳۵۱۰	۰/۷۴۴۳۶۷۳	۰/۷۴۲۱۵۵۱
۰/۷۶۸۳۲۶۵	۰/۷۵۹۶۱۶۳	۰/۷۵۲۰۴۰۸	۰/۷۴۵۸۴۴۹	۰/۷۳۷۷۱۴۳	۰/۷۳۲۸۴۹۰
۰/۷۵۳۵۵۹۲	۰/۷۴۶۱۳۰۶	۰/۷۴۱۲۷۳۵	۰/۷۳۱۶۷۳۵	۰/۷۲۸۴۷۳۵	۰/۷۲۷۳۵۵۱

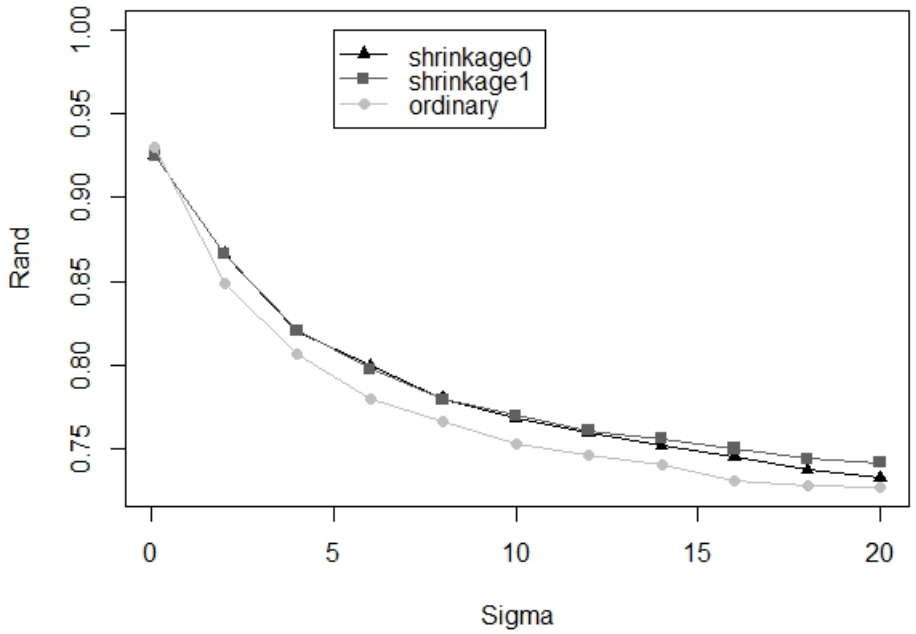


شکل ۱.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از دو زیر جامعه، بر اساس روش معمولی(دایره)، روش جیمز-استاین معمولی(مثلث) و جیمز-استاین بهینه(مربع)

هیچ‌کدام برتری قاطع بر دیگری ندارد ولی بطور کلی در این میان $b = ۱$ و $b = \hat{p} - ۲$ نسبت به بقیه‌ی bها در بیشتر مقادیر σ، عملکرد بهتری از خود نشان می‌دهند. با اختلاف بسیار کمی خوشه‌بندی مبتنی بر جیمزاستاین تعمیم‌یافته با $b = ۱$ بر همتای خود با $b = \hat{p} - ۲$ برتری دارد.

مثال ۳.۲.۳. در این مثال با جامعه‌ای متشکل از ۹ زیرجامعه با توزیع نرمال ۵ متغیره مواجه هستیم، ماتریس کواریانس جامعه چون مثال‌های قبل اختیار می‌شود، بردارهای میانگین نیز در برنامه‌نویسی مربوطه قابل رویت است. بنابر جدول ۳.۳ در میان مقادیر b، برآوردگر جیمزاستاین متناظر با $b = \hat{p} - ۲$ عملکرد دقیق‌تری در خوشه‌بندی دارد. همچنین نمودار ۳.۳ نشان می‌دهد، اگرچه خوشه‌بندی به روش جیمز-استاین معمولی برای واریانس درون خوشه‌ای بزرگتر از ۱۶ نسبت به K-میانگین سنتی افت می‌کند ولی روش جیمز-استاین تعمیم‌یافته با $b = \hat{p} - ۲$ و یا $b = ۱$ این افت را جبران می‌کند.

مثال ۴.۲.۳. در این مثال تعداد زیرجوامع و به دنبال آن خوشه‌ها تا ۱۳ افزایش می‌یابد. با توجه به

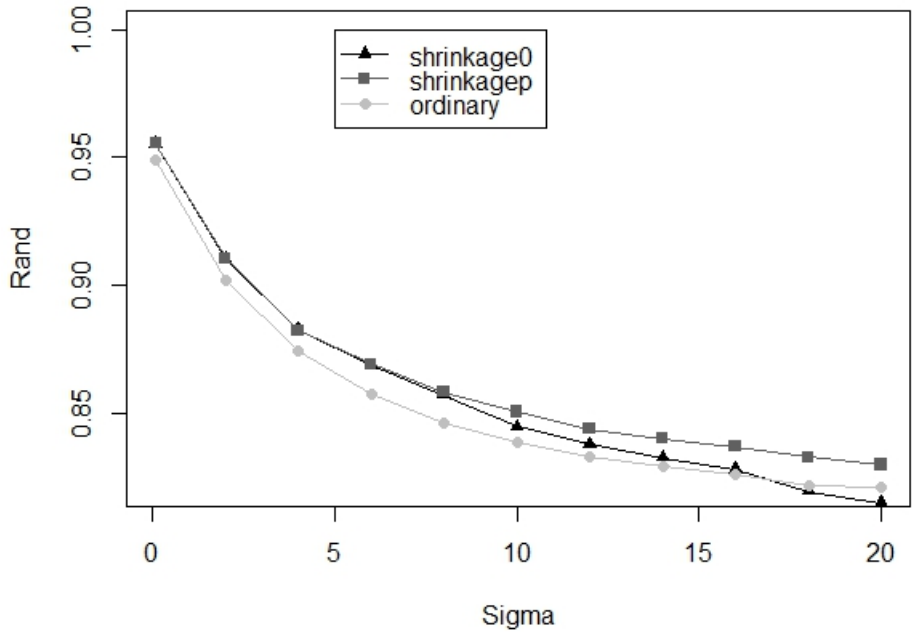


شکل ۲.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از پنج زیرجامعه در مثال ۲.۲.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین معمولی(مثلث) و جیمز-استاین بهینه(مربع)

جدول ۳.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای نه زیرجامعه در مثال ۳.۲.۳

روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$
تعمیم یافته	۰/۹۵۴۹۵۳۸	۰/۹۰۹۲۶۳۴	۰/۸۸۴۱۰۷۴	۰/۸۷۰۰۵۲۴	۰/۸۵۸۴۴۴۴
انقباضی	۰/۹۵۴۹۵۶۳	۰/۹۰۹۲۸۳۴	۰/۸۸۲۲۵۲۲	۰/۸۶۷۶۷۵۴	۰/۸۵۷۰۵۶۲
معمولی	۰/۹۴۵۸۸۲۶	۰/۹۰۳۳۰۵۹	۰/۸۷۳۷۴۲۸	۰/۸۵۷۲۰۶۰	۰/۸۴۴۶۸۴۱
$\sigma = ۱۰$	$\sigma = ۱۲$	$\sigma = ۱۴$	$\sigma = ۱۶$	$\sigma = ۱۸$	$\sigma = ۲۰$
۰/۸۴۹۴۴۳۲	۰/۸۴۴۶۵۶۷	۰/۸۳۸۱۰۷۴	۰/۸۳۴۶۰۹۲	۰/۸۳۲۳۷۹۵	۰/۸۲۸۰۰۵۰
۰/۸۴۶۹۴۶۳	۰/۸۳۷۴۰۳۲	۰/۸۳۲۶۸۱۶	۰/۸۲۶۲۴۷۲	۰/۸۲۱۱۷۶۰	۸۱۷۰۳۸۷
۰/۸۳۷۱۳۶۱	۰/۸۳۳۱۸۱۰	۰/۸۲۸۳۷۲۰	۰/۸۲۶۵۳۹۳	۰/۸۲۲۱۵۲۳	۰/۸۲۰۵۹۴۳

جدول ۴.۳ همان‌طور که از خروجی مقادیر مختلف b بر مقادیر متنوع σ استنباط می‌شود، تا $\sigma = ۱۲$، به‌طورکلی خوشه‌بندی به روش جیمز-استاین تعمیم یافته به ازای تمامی مقادیر b درنظرگرفته شده، نسبت



شکل ۳.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از نه زیر جامعه در مثال ۳.۲.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین معمولی(مثلث) و جیمز-استاین بهینه(مربع)

به K-میانگین سنتی برتر است و یا این‌که اختلافشان قابل چشم‌پوشی است. به ازای $\sigma \geq ۱۴$، جیمز-استاین معمولی، در مقابل K-میانگین سنتی به تدریج افت می‌کند، درحالی‌که جیمز-استاین تعمیم‌یافته به ازای سایر مقادیر b، همچنان در مقابل افزایش واریانس درون خوشه‌ای برتری خود را حفظ می‌کند. برترین برآوردگر جیمز-استاین تعمیم‌یافته، برآوردگر مبتنی بر $b = \hat{p} - ۲$ است، که همواره نتایج قابل قبولی دارد.

۳.۳ خوشه‌بندی با استفاده از برآوردگر بارانچیک

فرض کنید بردار $X = (X_۱, X_۲, ..., X_p)$ دارای توزیع نرمال p-متغیره با بردار میانگین $\theta = (\theta_۱, ..., \theta_p)$ و ماتریس کواریانس معین مثبت Q باشد. بارانچیک[۲] (۱۹۶۴) برآوردگر زیر را ارائه

[۲] Baranchik

جدول ۴.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای سیزده زیرجامعه در مثال ۴.۲.۳

روش	$\sigma = ۰/۱$	$\sigma = ۲$	$\sigma = ۴$	$\sigma = ۶$	$\sigma = ۸$
تعمیم یافته	۰/۹۵۹۹۵۱۱	۰/۹۰۹۲۶۳۴	۰/۸۸۴۱۰۷۴	۰/۸۷۰۰۵۲۴	۰/۸۵۸۴۴۴۴
انقباضی	۰/۹۵۹۹۵۱۱	۰/۹۳۳۳۹۵۳	۰/۹۱۲۷۹۳۱	۰/۹۰۰۲۴۸۱	۰/۸۹۱۶۳۲۷
معمولی	۰/۹۶۰۱۳۸۳	۰/۹۲۷۴۷۲۹	۰/۹۰۴۱۰۶۱	۰/۸۹۲۷۷۷۶	۰/۸۸۳۷۷۹۴
$\sigma = ۱۰$	$\sigma = ۱۲$	$\sigma = ۱۴$	$\sigma = ۱۶$	$\sigma = ۱۸$	$\sigma = ۲۰$
۰/۸۴۹۴۴۳۲	۰/۸۸۳۵۴۳۲	۰/۸۸۱۰۳۷۶	۰/۸۷۸۵۲۸۳	۰/۸۷۷۲۸۴۴	۰/۸۷۵۱۶۰۴
۰/۸۸۷۰۳۵۲	۰/۸۷۲۶۱۴۲	۰/۸۶۶۷۹۴۳	۰/۸۶۱۲۷۴۹	۰/۸۶۰۲۶۲۴	۸۵۱۹۱۴۱
۰/۸۷۹۹۷۱۴	۰/۸۷۵۸۷۱۲	۰/۸۷۳۳۹۴۲	۰/۸۷۰۱۸۸۴	۰/۸۶۸۸۹۶۸	۰/۸۶۸۰۹۹۰

داد.

$$\delta_B(X) = (۱ - \frac{(\hat{p} - ۲)r(F)}{F})X \qquad (۷.۳)$$

که $\hat{p} = \frac{tr(Q)}{\lambda_{\max}(Q)}$، $F = x^T Q^{-۱} x$ و r یک تابع نزولی و کراندار می‌باشد.

در صورتی‌که ضریب انقباض منفی باشد یک جنبه‌ی نامناسب از برآوردگر است به همین دلیل ما در مطالعات شبیه‌سازی خود از جزء مثبت برآوردگر بارانچیک استفاده می‌کنیم.

$$\delta_{B^+}(X) = (۱ - \frac{(\hat{p} - ۲)r(F)}{F})^+ X$$

با انتخاب $r = \frac{P}{P+F}$ واضح است که $۰ \leq r \leq ۱$ از طرفی

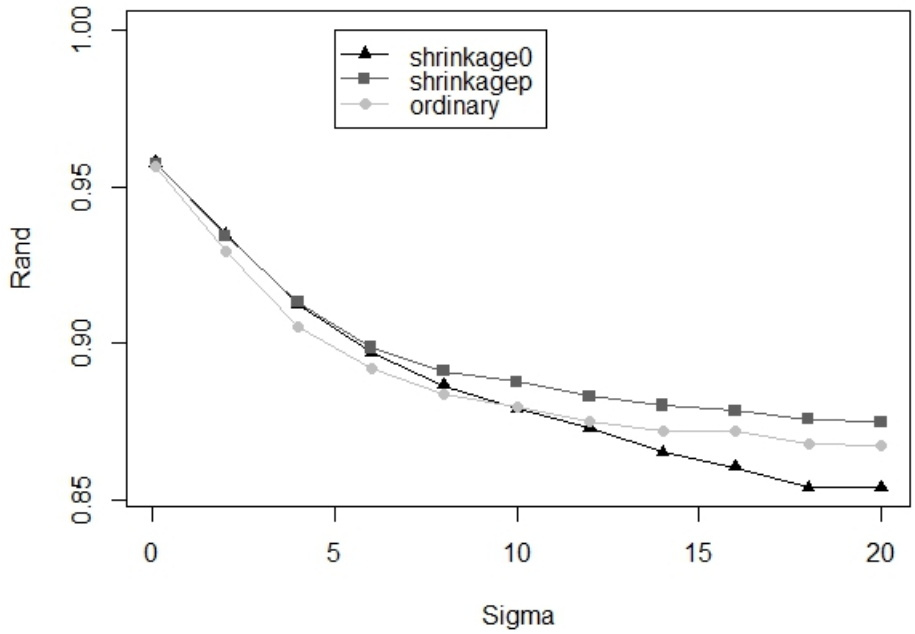
$$\frac{\partial r}{\partial F} = \frac{-P}{(F+P)^۲} \leq ۰$$

بنابراین با جایگذاری r در (۷.۳) برآوردگر بارانچیک به صورت زیر به‌دست می‌آید:

$$\delta_B(\bar{X}) = \bar{X} + \left[۱ - \frac{p\,(\hat{p} - ۲)}{F\,(p+F)}\right](\bar{X}_i - \bar{X})$$

نظیر آن‌چه در مورد برآوردگر جیمز-استاین معمولی و تعمیم‌یافته اتفاق افتاد، برآوردگر بارانچیک نیز قابل به‌کارگیری در انقباض مراکز خوشه‌ها به روش $K-$میانگین به سمت میانگین کل است. در نتیجه از رابطه‌ی زیر برای استفاده در الگوریتم خوشه‌بندی $K-$میانگین استفاده می‌کنیم:

$$\delta_B(\bar{X}) = \bar{X} + \left[۱ - \frac{p\,(\hat{p} - ۲)}{F\,(p+F)}\right]^+ (\bar{X}_i - \bar{X})$$



شکل ۴.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از سیزده زیرجامعه در مثال ۴.۲.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین معمولی(مثلث) و جیمز-استاین بهینه(مربع)

هم‌چون برآوردگر جیمزاستاین تعمیم‌یافته در قسمت قبل، عملکرد این برآوردگر و مقایسه‌ی آن با برآوردگرهای K-میانگین سنتی و جیمز-استاین معمولی در مثال‌های متنوعی مورد مطالعه و بررسی قرار می‌گیرند. این مثال‌ها، کاملا شبیه مثال‌های قبلی طرح شده می‌باشد، و تفاوت آن‌ها در برآوردگر مورد مطالعه است.

مثال ۱.۳.۳. (۲ خوشه‌ای)

جدول ۵.۳ و نمودار ۵.۳ تطبیق تقریبا کاملی را بین نتایج دو برآوردگر بارانچیک و جیمز-استاین نشان می‌دهد، همچنین هر دوی آن‌ها بر روش K-میانگین برتری محسوس دارند.

مثال ۲.۳.۳. (۵ خوشه‌ای)

در این مثال، روش بارانچیک در مقابل K-میانگین برتری تقریباً یکنواخت دارد. با توجه به جدول ۶.۳ در مقایسه‌ی عملکرد بارانچیک و جیمز-استاین معمولی، به ازای $\sigma \leq ۱۴$، بارانچیک برتری جزیی بر جیمز-استاین معمولی دارد، ولی به ازای سایر مقادیر σ، برتری بارانچیک بر جیمز-استاین معمولی

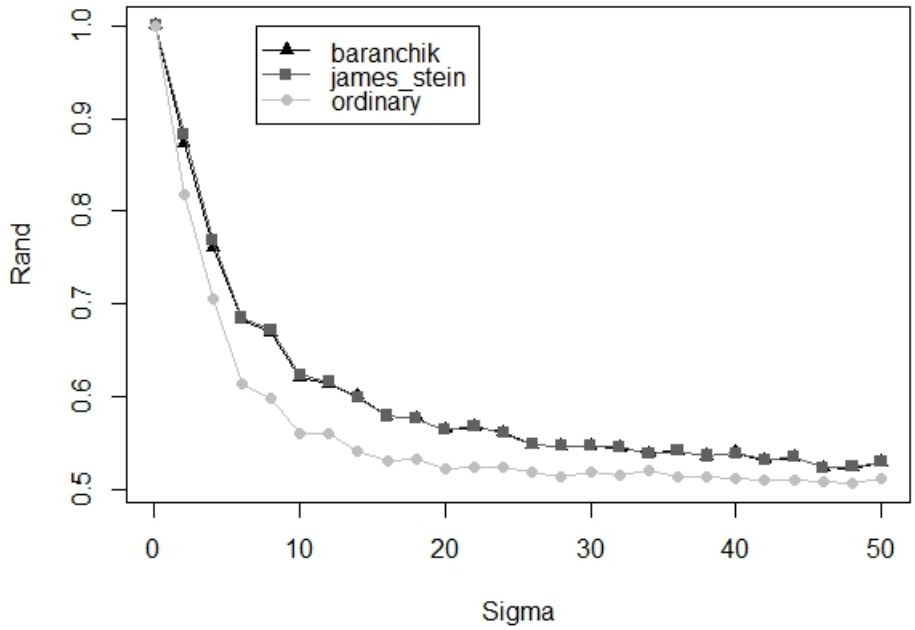
جدول ۵.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای دو زیرجامعه در مثال ۱.۳.۳

روش	$\sigma = ۰/۱$	$\sigma = ۴$	$\sigma = ۱۰$	$\sigma = ۱۴$
بارانچیک	۱/۰۰۰۰۰۰۰	۰/۷۶۰۸۴۹۰	۰/۶۲۱۳۹۵۹	۰/۶۰۱۱۳۴۷
انقباضی	۱/۰۰۰۰۰۰۰	۰/۷۶۸۴۶۵۳	۰/۶۲۴۲۵۳۱	۰/۵۹۹۱۸۳۷
معمولی	۱/۰۰۰۰۰۰۰	۰/۷۰۵۳۳۸۸	۰/۵۶۱۱۰۲۰	۰/۵۴۱۹۳۴۷
روش	$\sigma = ۲۰$	$\sigma = ۲۴$	$\sigma = ۳۰$	$\sigma = ۳۴$
بارانچیک	۰/۵۴۶۸۸۱۶	۰/۵۳۹۲۶۵۳	۰/۶۲۱۳۹۵۹	۰/۶۰۱۱۳۴۷
انقباضی	۰/۵۴۷۳۳۰۶	۰/۵۳۸۷۱۰۲	۰/۶۲۴۲۵۳۱	۰/۵۹۹۱۸۳۷
معمولی	۰/۵۱۸۵۰۶۱	۰/۵۲۰۰۲۴۵	۰/۵۶۱۱۰۲۰	۰/۵۴۱۹۳۴۷
روش	$\sigma = ۴۰$	$\sigma = ۴۴$	$\sigma = ۴۸$	$\sigma = ۵۰$
بارانچیک	۰/۵۴۰۳۳۴۷	۰/۵۳۴۸۹۸۰	۰/۵۲۳۲۶۵۳	۰/۵۲۹۴۲۰۴
انقباضی	۰/۵۳۹۴۲۸۶	۰/۵۳۵۴۸۵۷	۰/۵۲۴۸۱۶۳	۰/۵۳۰۲۴۴۹
معمولی	۰/۵۱۲۴۸۹۸	۰/۵۰۹۶۹۸۰	۰/۵۰۶۷۶۷۳	۰/۵۱۲۲۳۶۷

مشهود است، و در شرایطی که جیمز-استاین معمولی به ازای $\sigma > ۲۲$، مغلوب $K-$میانگین می‌شود، بارانچیک بطور یکنواخت برتری خود بر $K-$میانگین را حفظ می‌کند.

جدول ۶.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای پنج زیرجامعه برای مثال ۲.۳.۳

روش	$\sigma = ۰/۱$	$\sigma = ۴$	$\sigma = ۱۰$	$\sigma = ۱۴$
بارانچیک	۰/۹۳۰۶۰۴۱	۰/۸۲۰۲۸۵۷	۰/۷۷۲۸۲۴۵	۰/۷۵۸۵۴۶۹
انقباضی	۰/۹۳۱۶۴۰۸	۰/۸۱۹۹۴۲۹	۰/۷۷۱۶۸۱۶	۰/۷۵۱۷۴۶۹
معمولی	۰/۹۳۷۷۴۶۹	۰/۸۰۷۵۸۳۷	۰/۷۵۴۶۶۱۲	۰/۷۳۹۰۱۲۲
روش	$\sigma = ۲۰$	$\sigma = ۲۴$	$\sigma = ۳۰$	$\sigma = ۳۴$
بارانچیک	۰/۷۳۷۴۶۹۴	۰/۷۳۶۳۵۹۲	۰/۷۲۷۹۹۱۸	۰/۷۱۹۹۲۶۵
انقباضی	۰/۷۲۹۷۷۹۶	۰/۷۱۸۵۹۵۹	۰/۷۰۲۲۵۳۱	۰/۶۸۱۸۹۳۹
معمولی	۰/۷۲۵۵۷۵۵	۰/۷۲۰۳۳۴۷	۰/۷۱۶۷۷۵۵	۰/۷۰۵۷۳۰۶
روش	$\sigma = ۴۰$	$\sigma = ۴۴$	$\sigma = ۴۸$	$\sigma = ۵۰$
بارانچیک	۰/۷۱۹۳۷۱۴	۰/۷۱۸۷۷۵۵	۰/۷۱۰۸۳۲۷	۰/۷۱۲۳۵۹۲
انقباضی	۰/۶۸۱۲۸۱۶	۰/۶۷۲۳۲۶۵	۰/۶۶۰۱۷۹۶	۰/۶۶۵۱۸۳۷
معمولی	۰/۷۰۵۹۰۲۰	۰/۷۰۵۹۴۲۹	۰/۶۹۷۴۴۴۹	۰/۷۰۰۵۳۰۶



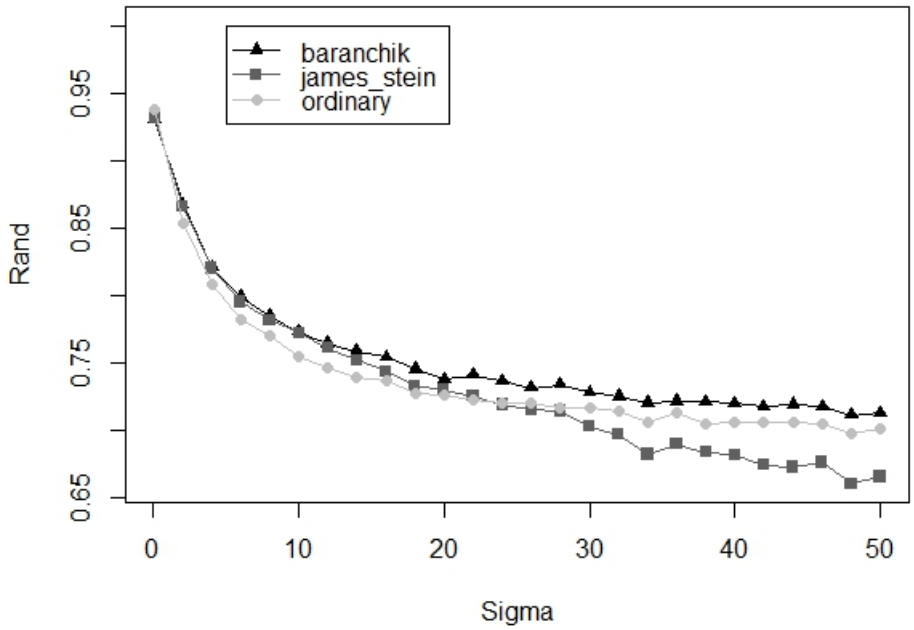
شکل ۵.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از دو زیرجامعه در مثال ۱.۳.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین معمولی(مربع) و بارانچیک(مثلث)

مثال ۳.۳.۳. (۹ خوشه‌ای)

با توجه به خروجی‌های جدول ۷.۳ و نمودار ۷.۳، این مثال الگوی مشابهی چون مثال قبل از خود نشان می‌دهد. به این‌صورت که روش بارانچیک، در برابر روش K-میانگین برتری کمی دارد، ولی به ازای $\sigma \geq ۳۰$، این برتری به حداقل خود می‌رسد. همچنین، به ازای $\sigma \leq ۱۴$، بارانچیک در برابر جیمز-استاین معمولی برتری ناچیزی دارد، ولی به‌ازای $\sigma > ۱۴$، رفته‌رفته، این برتری بطور معناداری افزایش می‌یابد، در حالی‌که روش جیمز-استاین معمولی از روش K-میانگین، به شدت عقب می‌افتد.

مثال ۴.۳.۳. (۱۳ خوشه‌ای)

خروجی‌های جدول ۸.۳ و نمودار ۸.۳ نشان می‌دهد، در رقابت بین این سه برآوردگر، به ازای $\sigma \leq ۸$، تقریبا هر سه دقت یکسانی دارند، ولی به‌ازای $\sigma > ۸$، افت جیمز-استاین معمولی شروع می‌شود و همانند مثال‌های قبلی پیوسته دقتش کاهش می‌یابد. در این مثال روش بارانچیک که به‌ازای $\sigma \leq ۳۴$



شکل ۶.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از پنج زیرجامعه در مثال ۲.۳.۳، بر اساس روش معمولی(دایره)، روش جیمز–استاین معمولی(مربع) و بارانچیک(مثلث)

برتری جزیی بر K–میانگین دارد، به‌ازای $\sigma > ۳۴$، این برتری جزیی را به K–میانگین می‌دهد.

از مثال‌های مطرح شده می‌توان نتیجه گرفت با افزایش تعداد خوشه‌ها، خوشه‌بندی به روش جیمز–استاین معمولی به ازای واریانس‌های درون خوشه‌ای نه چندان بزرگ نسبت به K–میانگین سنتی برتری دارد. به‌عبارت دیگر، جیمز–استاین معمولی در خوشه‌بندی داده‌هایی با واریانس درون خوشه‌ای خیلی بزرگ مناسب نیست. در این موارد برآوردگرهای مطرح شده در این فصل، یعنی برآوردگر جیمز استاین تعمیم‌یافته و بارانچیک، نسبت به افزایش σ مقاوم‌تر هستند، و در مقابل K–میانگین سنتی، برتری قابل قبولی دارند و در بدترین حالت، برآوردگر بارانچیک در مثال ۴.۳.۳، نسبت به K–میانگین سنتی افت محسوسی دارد، در عوض برآوردگر جیمز–استاین بهینه در این مثال برتری کامل دارد، که منظور از آن، برآوردگر جیمز–استاین تعمیم‌یافته به ازای b بهینه است.

مشکل نه چندان جدی که در مثال ۴.۳.۳، برای برآوردگر بارانچیک به‌وجود آمد، ما را ترغیب می‌کند که به مقایسه‌ی برآوردگر جیمز استاین بهینه با بارانچیک بپردازیم و در صورت امکان روشی را برگزینیم که

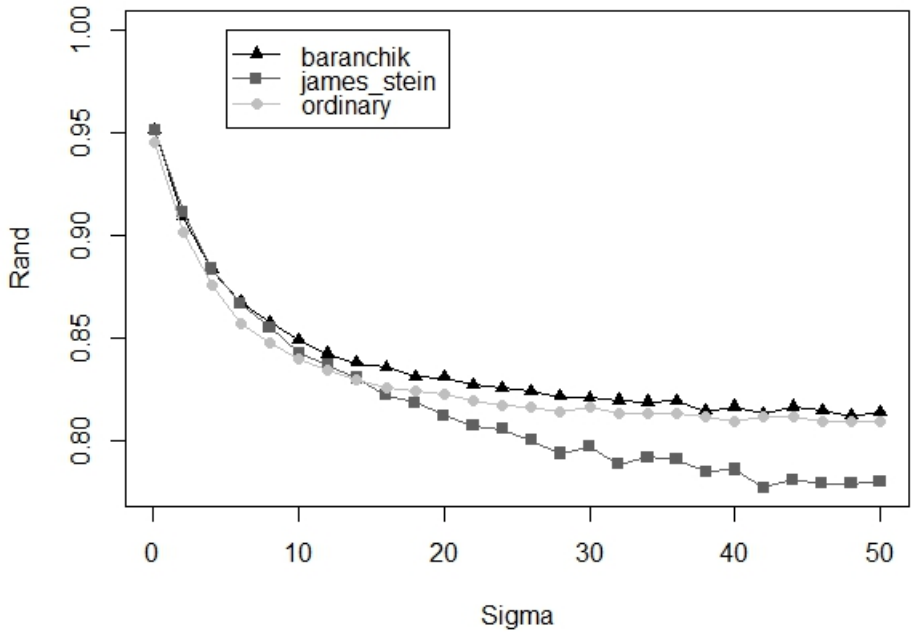
جدول ۷.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای نه زیرجامعه در مثال ۳.۳.۳

روش	$\sigma = ۰/۱$	$\sigma = ۴$	$\sigma = ۱۰$	$\sigma = ۱۴$
بارانچیک	۰/۹۵۱۱۰۳۶	۰/۸۸۳۰۵۸۷	۰/۸۴۸۹۳۶۳	۰/۸۳۷۹۷۰۰
انقباضی	۰/۹۵۱۰۷۱۲	۰/۸۸۳۶۰۰۵	۰/۸۴۲۸۲۶۵	۰/۸۳۰۶۶۶۷
معمولی	۰/۹۴۵۰۷۸۷	۰/۸۷۵۵۲۸۱	۰/۸۳۹۵۸۳۰	۰/۸۳۰۰۹۴۹
روش	$\sigma = ۲۰$	$\sigma = ۲۴$	$\sigma = ۳۰$	$\sigma = ۳۴$
بارانچیک	۰/۸۳۰۸۴۸۹	۰/۸۲۵۶۸۰۴	۰/۸۲۰۸۹۸۹	۰/۸۱۸۹۸۶۳
انقباضی	۰/۸۱۲۴۰۷۰	۰/۸۰۵۸۵۲۷	۰/۷۹۷۱۸۶۰	۰/۷۹۲۲۰۷۲
معمولی	۰/۸۲۲۷۲۴۱	۰/۸۱۶۹۸۸۸	۰/۸۱۶۴۷۴۴	۰/۸۱۳۷۶۰۳
روش	$\sigma = ۴۰$	$\sigma = ۴۴$	$\sigma = ۴۸$	$\sigma = ۵۰$
بارانچیک	۰/۸۱۶۷۷۹۰	۰/۸۱۶۵۷۱۸	۰/۸۱۲۱۴۹۸	۰/۸۱۴۰۲۲۵
انقباضی	۰/۷۸۶۱۴۹۸	۰/۷۸۱۳۴۵۸	۰/۷۷۹۳۷۵۸	۰/۷۸۰۳۸۲۰
معمولی	۰/۸۰۹۳۶۳۳	۰/۸۱۲۰۴۴۹	۰/۸۰۹۷۸۰۳	۰/۸۰۹۳۷۸۳

جدول ۸.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از سه روش برای سیزده زیرجامعه در مثال ۴.۳.۳

روش	$\sigma = ۰/۱$	$\sigma = ۴$	$\sigma = ۱۰$	$\sigma = ۱۴$
بارانچیک	۰/۹۶۱۹۵۱۱	۰/۹۱۱۰۸۷۷	۰/۸۸۵۸۷۷۲	۰/۸۷۸۶۱۰۶
انقباضی	۰/۹۶۱۹۲۶۱	۰/۹۱۱۱۳۰۶	۰/۸۷۷۰۳۷۶	۰/۸۶۳۶۶۹۶
معمولی	۰/۹۵۷۹۲۱۳	۰/۹۰۴۶۳۶۹	۰/۸۷۹۶۷۲۰	۰/۸۷۳۷۹۶۱
روش	$\sigma = ۲۰$	$\sigma = ۲۴$	$\sigma = ۳۰$	$\sigma = ۳۴$
بارانچیک	۰/۸۷۲۳۱۱۳	۰/۸۶۹۴۵۲۶	۰/۸۶۶۱۹۸۰	۰/۸۶۳۳۶۳۱
انقباضی	۰/۸۴۳۷۲۶۹	۰/۸۳۶۴۹۰۲	۰/۸۶۶۱۹۸۰	۰/۸۶۳۳۶۳۱
معمولی	۰/۸۶۶۹۹۲۲	۰/۸۶۶۸۰۸۶	۰/۸۶۴۷۰۱۳	۰/۸۶۳۱۸۴۳
روش	$\sigma = ۴۰$	$\sigma = ۴۴$	$\sigma = ۴۸$	$\sigma = ۵۰$
بارانچیک	۰/۸۶۱۷۲۲۱	۰/۸۶۵۸۹۰۳	۰/۸۵۶۱۶۸۲	۰/۸۵۸۳۳۲۷
انقباضی	۰/۸۱۹۴۲۹۹	۰/۸۱۱۱۵۳۲	۰/۸۱۳۲۷۷۳	۰/۸۱۱۲۴۲۷
معمولی	۰/۸۶۴۲۴۹۳	۰/۸۶۱۷۹۲۵	۰/۸۶۱۶۷۸۰	۰/۸۶۱۵۵۱۶

نتایج بهتری به‌بار آورد. مثال‌های قبلی را مجدداً برای مقایسه‌ی این دو برآوردگر به‌کار می‌گیریم. بدون بازگویی مثال‌ها، تنها به خروجی و تفسیر مربوطه در هر مثال می‌پردازیم.



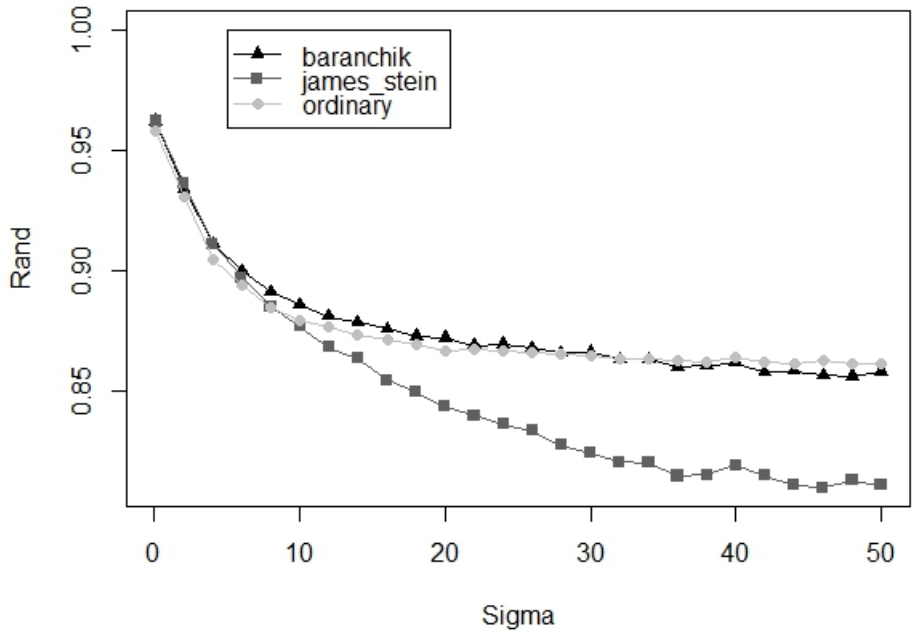
شکل ۷.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از نه زیرجامعه در مثال ۳.۳.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین معمولی(مربع) و بارانچیک(مثلث)

مثال ۵.۳.۳. (۵ خوشه)

در این مثال، برآوردگر جیمز-استاین تعمیم‌یافته در $\hat{p}-۲$ نسبتا بهینه است. جدول ۹.۳ و نمودار ۹.۳ نشان می‌دهد، تطبیق نسبتاً کاملی بین خوشه‌بندی به روش جیمز-استاین بهینه و بارانچیک وجود دارد. اختلاف شاخص رند در این دو روش معنادار نیست، با این حال برتری بسیار جزیی بارانچیک بر جیمز-استاین بهینه، به‌ازای $\sigma < ۲۸$ وجود دارد، و به‌ازای $\sigma \leq ۲۸$ این برتری بسیار جزیی در جیمز-استاین بهینه نسبت به بارانچیک وجود دارد.

مثال ۶.۳.۳. (۹ خوشه)

در این مثال طبق جدول ۱۰.۳ روش جیمز-استاین بهینه بر بارانچیک همواره غالب است. با این حال اختلاف شاخص رند این دو روش، در میان اندازه‌های مختلف واریانس درون خوشه‌ای هیچ‌گاه از ۰/۰۱



شکل ۸.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از سیزده زیرجامعه در مثال ۴.۳.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین معمولی(مربع) و بارانچیک(مثلث)

تجاوز نمی‌کند.

مثال ۷.۳.۳. (۱۳ خوشه)

این مثال به‌خوبی برتری روش جیمز-استاین بهینه را نشان می‌دهد. مشابه آن‌چه در مثال ۴.۳.۳ بیان شد، برآوردگر بارانچیک به ازای $\sigma \geq ۲۸$ به میزان ناچیزی در برابر K-میانگین سنتی افت می‌کند. در صورتی‌که براساس جدول ۱۱.۳ و نمودار ۱۱.۳، روش جیمز-استاین بهینه بطور معناداری نسبت به بارانچیک و K-میانگین سنتی برتری دارد، در برابر افزایش واریانس خوشه‌ای به‌خوبی مقاوم است و حتی نمودار شهودی نشان می‌دهد به‌ازای $\sigma \geq ۱۶$ مقدار شاخص رند بطور نامحسوسی کاهش می‌یابد.

از آن‌جا که در مثال ۴.۳.۳، که تعداد خوشه‌ها در آن ۲ می‌باشد، جیمز-استاین بهینه همان جیمز-استاین معمولی است. برای مطالعه‌ی خروجی‌ها و تفسیر مربوطه، به مثال نام‌برده می توان مراجعه کرد.

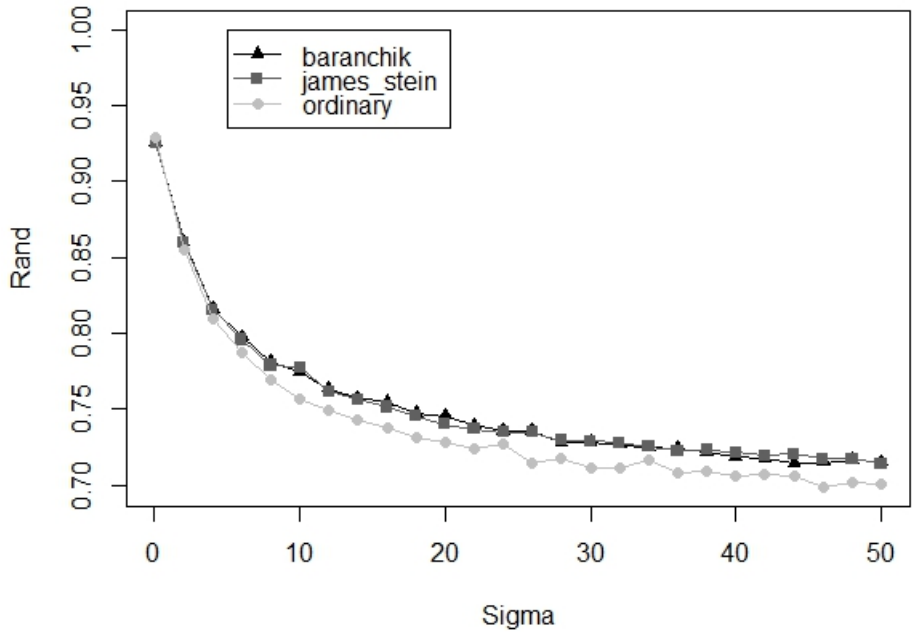
جدول ۹.۳: متوسط مقادیر شاخص رند برای مقایسه دو برآوردگر در خوشه‌بندی $K-$میانگین از پنج زیرجامعه در مثال ۵.۳.۳

روش	$\sigma = ۰/۱$	$\sigma = ۴$	$\sigma = ۱۰$	$\sigma = ۱۴$
بارانچیک	۰/۹۲۵۱۷۵۵	۰/۸۱۶۰۴۹۰	۰/۷۷۴۶۱۲۲	۰/۷۵۷۳۷۱۴
تعمیم یافته	۰/۹۲۵۱۸۳۷	۰/۸۱۵۵۱۸۴	۰/۷۷۷۳۳۸۸	۰/۷۵۶۳۴۲۹
معمولی	۰/۹۲۸۳۴۲۹	۰/۸۰۹۲۵۷۱	۰/۷۵۶۳۱۰۲	۰/۷۴۲۴۵۷۱
روش	$\sigma = ۲۰$	$\sigma = ۲۴$	$\sigma = ۳۰$	$\sigma = ۳۴$
بارانچیک	۰/۷۴۰۱۲۲۴	۰/۷۳۴۸۸۹۸	۰/۷۲۸۱۲۲۴	۰/۷۲۴۴۰۸۲
تعمیم یافته	۰/۷۴۰۱۲۲۴	۰/۷۳۴۸۸۹۸	۰/۷۲۹۰۰۴۱	۰/۷۲۵۴۳۶۷
معمولی	۰/۷۲۸۱۲۲۴	۰/۷۲۷۰۶۹۴	۰/۷۱۰۷۱۰۲	۰/۷۱۶۷۵۱۰
روش	$\sigma = ۴۰$	$\sigma = ۴۴$	$\sigma = ۴۸$	$\sigma = ۵۰$
بارانچیک	۰/۷۱۸۲۲۸۶	۰/۷۱۴۳۴۲۹	۰/۷۱۶۵۳۸۸	۰/۷۱۴۹۳۰۶
تعمیم یافته	۰/۷۲۱۵۵۱۰	۰/۷۲۰۴۳۲۷	۰/۷۱۷۰۷۷۶	۰/۷۱۳۹۴۲۹
معمولی	۰/۷۰۶۴۰۸۲	۰/۷۰۵۹۹۱۸	۰/۷۰۱۵۴۲۹	۰/۷۰۱۰۲۸۶

جدول ۱۰.۳: متوسط مقادیر شاخص رند برای مقایسه دو برآوردگر در خوشه‌بندی $K-$میانگین از نه زیرجامعه در مثال ۶.۳.۳

روش	$\sigma = ۰/۱$	$\sigma = ۴$	$\sigma = ۱۰$	$\sigma = ۱۴$
بارانچیک	۰/۹۵۱۶۴۵۴	۰/۸۸۳۰۶۳۷	۰/۸۴۸۶۱۹۲	۰/۸۳۷۶۰۸۰
تعمیم یافته	۰/۹۵۱۶۷۵۴	۰/۸۸۵۳۹۳۳	۰/۸۵۰۰۹۴۹	۰/۸۳۸۲۰۲۲
معمولی	۰/۹۴۹۳۳۳۳	۰/۸۷۴۶۳۶۷	۰/۸۳۹۷۶۵۳	۰/۸۲۶۵۸۹۳
روش	$\sigma = ۲۰$	$\sigma = ۲۴$	$\sigma = ۳۰$	$\sigma = ۳۴$
بارانچیک	۰/۸۲۸۶۰۱۷	۰/۸۲۵۷۵۰۳	۰/۸۲۰۱۱۹۹	۰/۸۱۸۲۵۹۷
تعمیم یافته	۰/۸۳۲۹۳۳۸	۰/۸۲۸۴۲۹۵	۰/۸۲۴۱۰۹۹	۰/۸۲۲۶۵۴۲
معمولی	۰/۸۲۱۴۳۸۲	۰/۸۱۷۶۸۲۹	۰/۸۱۴۱۳۴۸	۰/۸۱۴۳۷۹۵
روش	$\sigma = ۴۰$	$\sigma = ۴۴$	$\sigma = ۴۸$	$\sigma = ۵۰$
بارانچیک	۰/۸۱۶۱۱۷۴	۰/۸۱۳۳۲۸۳	۰/۸۱۱۸۱۰۲	۰/۸۱۳۰۳۶۲
تعمیم یافته	۰/۸۲۱۳۰۸۴	۰/۸۲۰۸۱۴۰	۰/۸۱۸۸۳۴۰	۰/۸۱۹۹۷۵۰
معمولی	۰/۸۱۰۹۲۸۸	۰/۸۱۰۴۳۷۰	۰/۸۰۹۴۹۸۱	۰/۸۱۰۲۱۲۲

نتایج مثال‌های فوق نشان می‌دهد، روش جیمز-استاین بهینه، بر روش بارانچیک برتری دارد، هرچند در بعضی موارد این برتری چندان محسوس نیست، ولی به‌هرحال بهترین عملکرد را دارد.

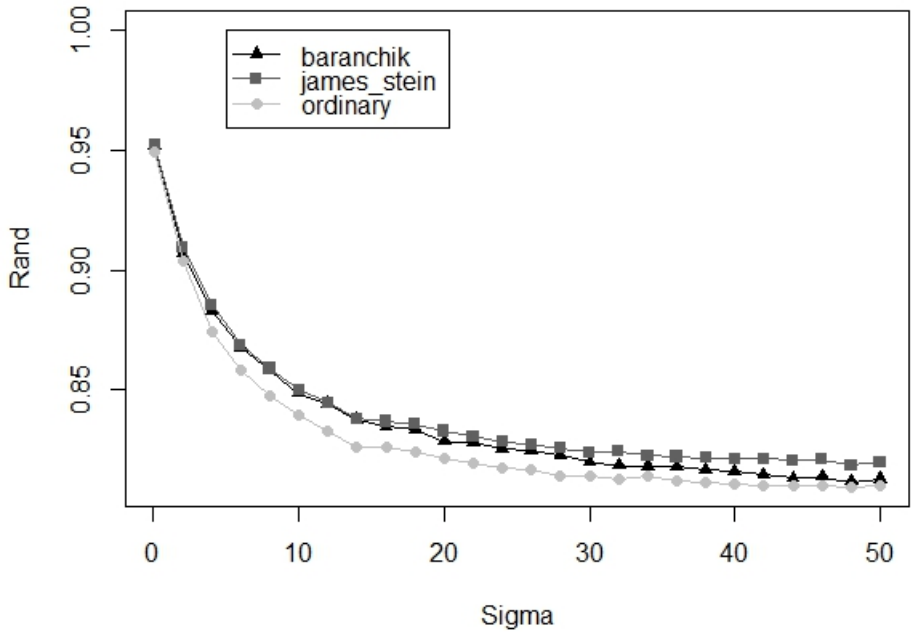


شکل ۹.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از پنج زیرجامعه در مثال ۵.۳.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین تعمیم‌یافته(مربع) و بارانچیک(مثلث)

۴.۳ خوشه‌بندی داده‌های سفال رومانو-بریتیش

در این بخش مجموعه داده‌های حاصل از تجزیه شیمیایی روی سفال‌های رومانو-بریتیش که در سه ناحیه ساخته شده را مورد بررسی قرار می‌دهیم. ناحیه‌ی اول شامل کوره‌ی شماره‌ی ۱، ناحیه‌ی دوم شامل کوره‌ی ۲ و ۳ و ناحیه‌ی سوم شامل کوره‌ی ۴ و ۵ است. مجموعه‌ی کامل داده‌ها شامل نتایج تجزیه‌ی شیمیایی ۴۵ کوزه میباشد که در اوریت و هوتورن (۲۰۱۱) آمده است.

ما با محاسبه‌ی ماتریس فاصله‌ی اقلیدسی برای مقادیر استانداردشده از ۴۵ کوزه، کار را شروع می‌کنیم که نتیجه یک ماتریس ۴۵×۴۵ می‌باشد و با استفاده از یک نمودار گرافیکی آن را بررسی می‌کنیم. این نمودار با تابع $levelplot$ که در بسته‌ی $lattice$ در دسترس است رسم شده است. کدهای مربوط به رسم نمودار به‌صورت زیر است.



شکل ۱۰.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از نه زیرجامعه در مثال ۶.۳.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین تعمیم‌یافته(مربع) و بارانچیک(مثلث)

```
pottery_dist <- dist(pots <- scale(pottery[, colnames(pottery) != "kiln"],
 center = FALSE))
library("lattice")
levelplot(as.matrix(pottery_dist), xlab = "Pot Number", ylab = "Pot Number")
```

این قبیل نمودارها هر خانه از ماتریس عدم مشابهت را با یک رنگ یا درجات مختلفی از رنگ خاکستری مرتبط می‌سازد (کد بالا مربوط به یک نمودار رنگی است). ما رنگ خاکستری بسیار تیره را برای خانه‌هایی با فاصله‌ی صفر (برای مثال عناصر قطر اصلی از ماتریس عدم مشابهت) و خاکستری کم‌رنگ برای خانه‌هایی با فاصله‌ی اقلیدسی بزرگ‌تر انتخاب کردیم. دستور آن به‌صورت زیر است.

```
trellis.par.set(standard.theme(color = FALSE))
plot(levelplot(as.matrix(pottery_dist), xlab = "Pot Number", ylab =
```

جدول ۱۱.۳: متوسط مقادیر شاخص رند برای مقایسه دو برآوردگر در خوشه‌بندی $K-$میانگین از سیزده زیرجامعه در مثال ۷.۳.۳

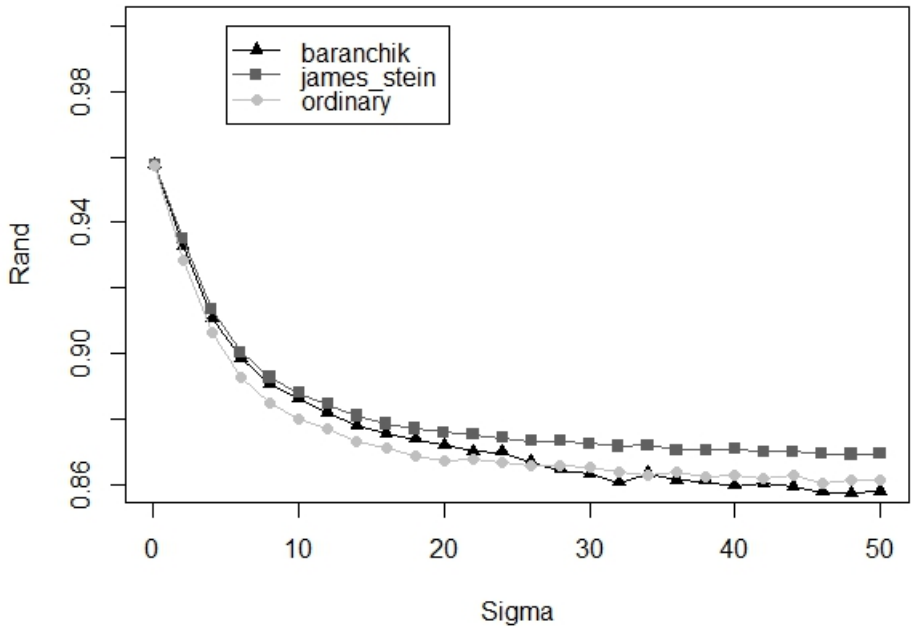
روش	$\sigma = ۰/۱$	$\sigma = ۴$	$\sigma = ۱۰$	$\sigma = ۱۴$
بارانچیک	۰/۹۵۷۵۶۷۱	۰/۹۱۰۸۳۱۲	۰/۸۸۶۴۷۳۵	۰/۸۷۸۲۹۸۲
تعمیم یافته	۰/۹۵۷۷۰۵۴	۰/۹۱۳۵۰۱۵	۰/۸۸۸۱۳۰۰	۰/۸۸۱۱۴۷۳
معمولی	۰/۹۵۷۱۵۹۲	۰/۹۰۶۳۵۵۴	۰/۸۸۰۰۴۵۳	۰/۸۷۳۰۴۸۳
روش	$\sigma = ۲۰$	$\sigma = ۲۴$	$\sigma = ۳۰$	$\sigma = ۳۴$
بارانچیک	۰/۸۷۲۲۴۰۹	۰/۸۷۰۰۲۹۸	۰/۸۶۳۳۳۶۹	۰/۸۶۳۵۵۴۰
تعمیم یافته	۰/۸۷۶۰۲۵۰	۰/۸۷۴۴۶۸۷	۰/۸۷۲۷۲۵۱	۰/۸۷۲۱۷۲۹
معمولی	۰/۸۶۷۲۴۷۵	۰/۸۶۶۷۴۶۶	۰/۸۶۵۳۱۷۸	۰/۸۶۳۰۷۸۱
روش	$\sigma = ۴۰$	$\sigma = ۴۴$	$\sigma = ۴۸$	$\sigma = ۵۰$
بارانچیک	۰/۸۵۹۷۷۸۲	۰/۸۵۹۷۰۶۶	۰/۸۵۷۴۸۲۴	۰/۸۵۸۲۵۶۴
تعمیم یافته	۰/۸۷۰۹۳۹۸	۰/۸۷۰۲۸۹۸	۰/۸۶۹۳۵۸۴	۰/۸۶۹۵۱۸۲
معمولی	۰/۸۶۲۸۲۵۳	۰/۸۶۲۹۰۱۶	۰/۸۶۱۷۴۷۲	۰/۸۶۱۵۶۳۵

```
"Pot Number", scales = list(x = list(draw = FALSE), y = list(draw = FALSE))))
```

طبق نمودار ۱۲.۳ مشخص می‌شود که حداقل سه گروه مجزا با اختلافات درون‌خوشه‌ای اندک (مربع تیره) وجود دارد در حالی‌که فاصله‌های خیلی بزرگتری می تواند برای خانه‌های دیگر مشاهده شود. برای مشاهده‌ی تأثیر تعداد خوشه‌ها در روش K-میانگین مجموع توان‌های دوم درون‌گروهی را برای یک تا شش گروه با استفاده از کدهای زیر رسم می‌کنیم.

```
n <- nrow(pots)

wss <- rep(0, 6)

wss[1] <- (n - 1) * sum(sapply(pots, var))

for (i in 2:6) wss[i] <- sum(kmeans(pots, centers = i)$withinss)

plot(1:6, wss, type = "b", xlab = "Number of groups",

ylab = "Within groups sum of squares")
```

مجدداً نمودار ۱۳.۳ به‌طور نسبتاً روشن نتیجه می‌دهد که داده‌ها شامل سه خوشه می‌باشند. اکنون

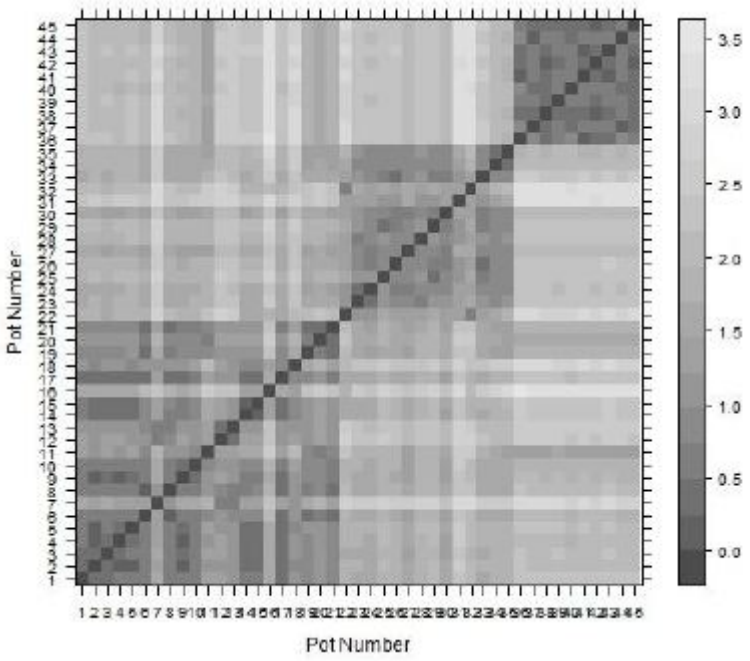


شکل ۱۱.۳: متوسط مقادیر شاخص رند برای خوشه‌بندی $K-$میانگین از ۵۰۰۰ داده ناهمبسته پنج بعدی از سیزده زیرجامعه در مثال ۷.۳.۳، بر اساس روش معمولی(دایره)، روش جیمز-استاین تعمیم‌یافته(مربع) و بارانچیک(مثلث)

مقایسه‌ای بین کوره‌هایی که سفال‌ها در آن قالب‌ریزی شده انجام می‌دهیم.

```
set.seed(29)

pottery_cluster <- kmeans(pots, centers = 3)$cluster

xtabs(~ pottery_cluster + kiln, data = pottery)
```

```
                    klin
pottery_cluster      1  2  3  4  5
              1     21  0  0  0  0
              2      0 12  2  0  0
              3      0  0  0  5  5
```

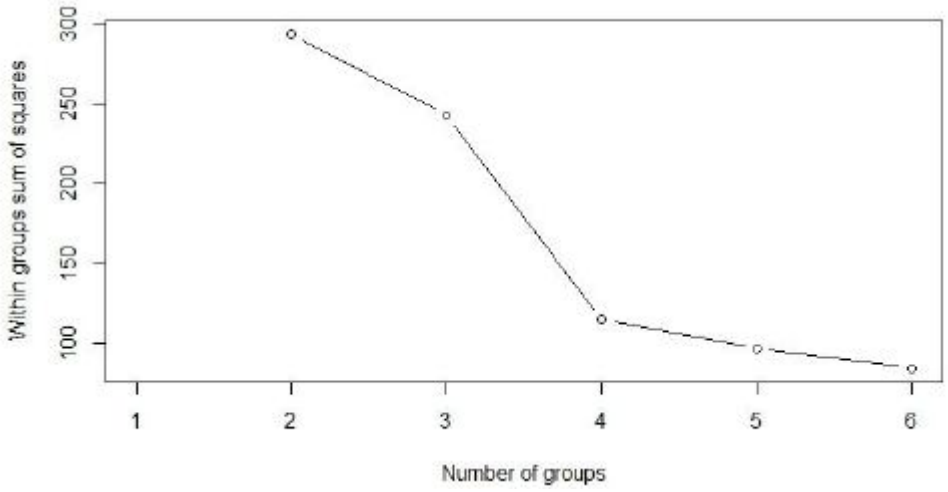


شکل ۱۲.۳: نمودار ماتریس عدم مشابهت داده‌های کوزه

جداول توافقی نشان می‌دهد که خوشه‌ی ۱ شامل همه‌ی کوزه‌های تولیدشده در کوره‌ی ۱ می‌باشد، خوشه‌ی ۲ شامل همه‌ی کوزه‌های تولیدشده در کوره‌های ۲ و ۳ و خوشه‌ی ۳ کوزه‌های تولیدشده در کوره‌ی ۴ و ۵ می‌باشد. درحقیقت مکان ۵ کوره در سه ناحیه هستند، ناحیه‌ی ۱ شامل فقط یک کوره، ناحیه ۲ شامل کوره‌های ۲ و ۳ و ناحیه‌ی ۳ شامل کوره‌های ۴ و ۵ می‌باشد. بنابراین خوشه های تشکیل شده درواقع به کوزه‌هایی از سه ناحیه مختلف ارتباط دارد. با فرض وجود سه گروه در فضای دو مؤلفه‌ی اول اصلی ماتریس همبستگی داده‌ها در نمودار ۱۴.۳ با استفاده از کدهای زیر نمایش داده شده است.

```
pots_pca <- prcomp(pots)
plot(pots_pca$x[, 1:2], pch = kmeans(pots, centers = 3)$cluster)
```

حال از دو برآوردگر ارائه‌شده در این فصل برای خوشه‌بندی داده‌های کوزه استفاده می‌کنیم و نتایج را با روش K-میانگین معمولی مقایسه می‌نماییم تا توانایی تفکیک خوشه‌ها را در روش جدید ارزیابی کنیم. داده‌های کوزه مربوط به ۵ کوره در ۳ ناحیه می‌باشند و همانطور که در بالا بررسی گردید بهترین گروه‌بندی برای داده‌ها، افراز آن‌ها به ۳ خوشه می‌باشد و این سه خوشه درواقع همان سه ناحیه هستند که کوره‌ها در آنجا واقع شده‌اند. برای داده‌های واقعی در این مثال نیز، $\hat{Q}_i$ به عنوان ماتریس کواریانس درون‌خوشه‌ای



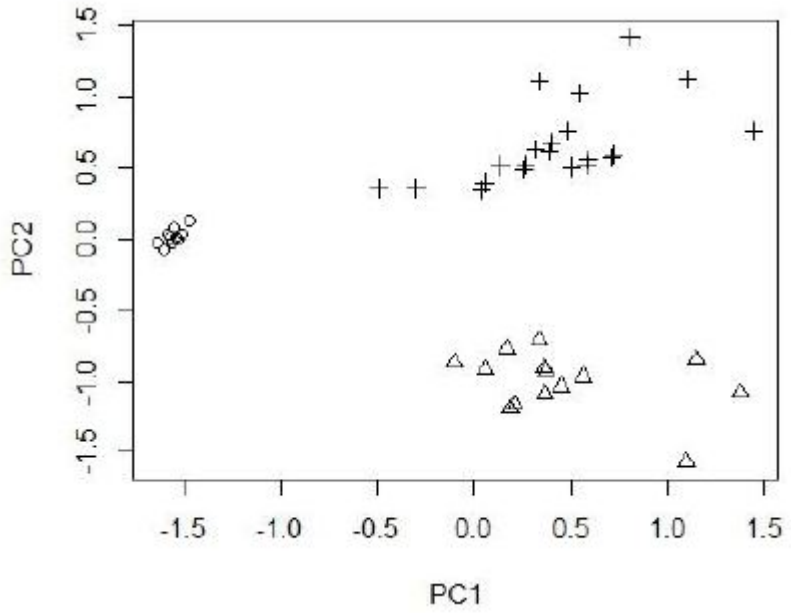
شکل ۱۳.۳: نمودار مجموع مربعات درون‌گروهی در مقابل تعداد خوشه‌ها

نمونه به‌جای Q_i استفاده می‌شود. براساس جدول ۱۲.۳، شاخص رَند در این مثال واقعی افزایش قابل

جدول ۱۲.۳: مقادیر شاخص رند برای داده‌های کوزه با استفاده از روش انقباضی و معمولی

روش	شاخص رند
بارانچیک	۰٫۸۲۴۳۴۱۱
جیمز-استاین تعمیم یافته	۰٫۸۳۱۲۱۲۱
معمولی	۰٫۷۷۵۷۵۷۶

توجهی نسبت به مثال واقعی مربوط به داده‌های ژن‌ها در فصل دوم پیدا کرده است که دلیل آن تفکیک مکانی کوزه‌ها از یکدیگر در مقابل ژن‌ها می‌باشد. همانطور که در جدول ۱۲.۳ مشاهده می‌شود دو برآوردگر بارانچیک و جیمز-استاین تعمیم‌یافته دارای شاخص رَند بالاتری نسبت به روش K-میانگین معمولی دارا می‌باشد. در برآوردگر جیمز-استاین تعمیم‌یافته از مقدار بهینه‌ی b یعنی $\hat{p}-۲$ استفاده شده و مانند نتایج قبل در داده‌های شبیه‌سازی‌شده دقت بیشتری دربرابر برآوردگر بارانچیک به‌دست آورده است.



شکل ۱۴.۳: نمودار K-میانگین سه‌گروهه برای داده‌های کوزه در فضای اولین زوج مؤلفه‌های اصلی از ماتریس همبستگی داده‌ها

۵.۳ نتیجه‌گیری

روش K-میانگین از دیرباز به‌عنوان روشی شناخته‌شده جهت خوشه‌بندی و هم‌چنین به‌عنوان روشی پایه جهت مقایسه با یک روش خوشه‌بندی جدیدتر مورد استفاده‌ی فراوان بوده است. با پیشرفت علوم مختلف، علاوه بر به‌وجود آمدن روش‌های خوشه‌بندی متفاوت با K-میانگین، خود الگوریتم K-میانگین دستخوش تغییرات متنوعی در ساختار شد، و حتی با سایر روش‌ها ادغام گردید. در این میان یکی از مفیدترین تغییرات ساختاری به‌وجود آمده در الگوریتم K-میانگین، منقبض نمودن مراکز خوشه‌ها به سمت میانگین کل با هر تکرار الگوریتم K-میانگین است. این تغییر ساختار باعث بهبود عملکرد الگوریتم K-میانگین گشت، ولی این بهبود در مسایل مختلف ثابت نیست و دچار افت و خیز می‌شود. از عواملی که نشان دادیم در حضور آن‌ها، روش جیمز-استاین بهبود زیادی را بوجود می‌آورد، عبارتند از:

۱. ناهمبستگی داده‌ها، به این معنا که میان توزیع هر بعد و دیگر بعدها ناهمبستگی وجود دارد.

۲. متوسط تا بزرگ بودن واریانس درون خوشه‌ای

۳. تفکیک مناسب بردار میانگین زیرجوامع از هم که موجب بزرگ بودن واریانس درون خوشه‌ای

است.

۴. داده‌های با ابعاد بالا

عکس این شرایط سبب کاهش بهبود جیمز-استاین می‌شود و حتی ممکن است $K-$میانگین در مقابل جیمز-استاین به برتری برسد.

در این‌جا سؤالی مطرح می‌شود؛ در مواردی که واریانس درون خوشه‌ای بسیار بزرگ است و به‌دنبال بهبود دقت خوشه‌بندی $K-$میانگین هستیم چه می‌توان کرد؟

در شرایط این چنینی که روش جیمز-استاین دچار افت می‌شود، می‌توان از روش‌های دیگری چون جیمز-استاین تعمیم‌یافته و بارانچیک بهره گرفت. البته جیمز-استاین تعمیم‌یافته به ازای مقادیر معینی از پارامتر b نتایج بهتری از خود نشان می‌دهد. این مقادیر بهینه را می‌توان از طریق آزمون و خطا یافت. نتایج حاصل از شبیه‌سازی مونت کارلو نشان می‌دهد به‌ازای تعداد خوشه‌های ۵ به بالا مقدار بهینه‌ی b که بیشترین دقت را سبب می‌شود، $\hat{p}-۲$ است. روش بارانچیک نیز در بیشتر مواقع خروجی‌های بهتری در مقایسه با روش‌های $K-$میانگین سنتی و جیمز-استاین معمولی از خود نشان می‌دهد، ولی با این حال نسبت به روش جیمز-استاین بهینه دقت کمتری دارد. بنابراین در مواردی که واریانس درون خوشه‌ای بسیار بزرگ است، روش جیمز-استاین بهینه با مقاومت در برابر افزایش واریانس درون‌خوشه‌ای بیشترین دقت را در خوشه‌بندی این بخش از آن خود کرده است.

در این‌جا باید به این نکته توجه کرد که لزومی ندارد از روش بارانچیک و به‌ویژه جیمز-استاین بهینه فقط در زمانی استفاده گردد که عوامل فوق نقص می‌شود. می‌توان تمامی مثال‌های موجود در برآوردگر جیمز-استاین معمولی را برای این دو برآوردگر محاسبه کرد.

۶.۳ پیشنهادات برای آینده تحقیق

با توجه به نتایج به‌دست آمده در این پایان‌نامه و موضوعات مطرح‌شده، می‌توان در آینده بر روی موارد زیر تحقیق کرد.

۱. مطالعه و بررسی برآوردگرهای انقباضی دیگر و همچنین توابع $c(.)$ به‌کار رفته در برآوردگر بارانچیک تا با بهره‌گیری از آنها بتوان به روش‌های جدید برای بهبود بیشتر روش K-میانگین و خوشه‌بندی داده‌ها دست یافت.

۲. تکنیک‌های مطرح‌شده برای خوشه‌بندی را در توزیع‌های دیگر از قبیل t-استیودنت چندمتغیره به‌جای نرمال چندمتغیره به‌کار گرفت.

۳. برای مقایسه‌ی عملکرد روش‌های خوشه‌بندی از شاخص‌های دیگر از قبیل شاخص رند تعدیل‌شده و شاخص نیمرخ استفاده کرد.

۴. روش K-میانگین شامل الگوریتم‌های مختلفی می‌باشد که رایج‌ترین آن مربوط به مک کوئین (۱۹۶۷) می‌باشد، مطالعات مشابهی می توان با الگوریتم فورجی[۳] (۱۹۶۵) یا الگوریتم‌های دیگر نیز انجام داد.

۵. رهیافت خوشه‌بندی افرازی شامل روش‌های پرکاربردی مانند K-مدوید می‌باشد، می‌توان تکنیک‌های مطرح‌شده در این پایان‌نامه را برای روش K-مدوید نیز به‌کار برد.

[۳] Forgy

پیوست آ

تعاریف

در این بخش تعاریف اساسی که در طول این مجموعه از آن‌ها استفاده شده بیان می‌شود.

تعریف آ.۰.۱. نرم اقلیدسی

$\|x\|$ را روی بردار $x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ نرم اقلیدسی گوییم هرگاه

$$\|x\| = (x^T x)^{\frac{1}{2}} = (\sum_{i=1}^{p} x_i^2)^{1/2}.$$

تعریف آ.۰.۲. تابع زیان توان دوم

اگر بردار p مولفه‌ای $\delta = (\delta_1, \ldots, \delta_p)$ برآوردگری برای بردار پارامتر $\theta \in \mathbb{R}^p$ باشد، آنگاه زیان استفاده از δ در خصوص برآورد θ را با $L(\theta, \delta)$ نشان داده که برابر است با

$$L(\theta, \delta) = (\delta - \theta)^T(\delta - \theta) = \|\delta - \theta\|^2 = \sum_{i=1}^{p}(\delta_i - \theta_i)^2,$$

چنان‌چه پارامتر مقیاس مجهول $\sigma^2 \geq 0$ در مدل موجود باشد، آنگاه از تابع زیان پایا زیر استفاده می‌کنیم.

$$L(\theta, \delta) = \frac{(\delta - \theta)^T(\delta - \theta)}{\sigma^2} = \frac{\|\delta - \theta\|^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{p}(\delta_i - \theta_i)^2.$$

تعریف آ.۰.۳. تابع مخاطره

با استفاده از تعریف تابع زیان، تابع مخاطره برآوردگر δ را با $R(\theta, \delta)$ نشان داده که برابر است با

$$R(\theta, \delta) = E_\theta[L(\theta, \delta)] = \int_\chi L(\theta, \delta) f_\theta(x) dx.$$

که در آن $f_\theta(\cdot)$ چگالی تعریف شده روی χ است.

تعریف آ.۴.۰. برآوردگر برتر

فرض کنید δ_1 و δ_2 دو برآوردگر $g(\theta)$ باشند. اگر به ازای هر $\theta \in \Theta$، داشته باشیم

$$R(\theta, \delta_1) \leq R(\theta, \delta_2),$$

و به ازای حداقل یک $\theta \in \Theta$ نامساوی اکید باشد، آنگاه برآوردگر δ_1 را بهتر (برتر) از برآوردگر δ_2 گوییم. به عبارت دیگر گوییم δ_1 اکیداً بر δ_2 غلبه دارد.

تعریف آ.۵.۰. برآوردگر غیرمجاز

برآوردگر δ را غیرمجاز گوییم هرگاه برآوردگر δ^* وجود داشته باشد به طوری که به ازای هر $\theta \in \Theta$ داشته باشیم

$$R(\theta, \delta^*) \leq R(\theta, \delta),$$

و حداقل یک مقدار $\theta \in \Theta$ وجود داشته باشد به طوری که نامساوی اکید باشد. هرگاه چنین برآوردگری وجود نداشته باشد، آنگاه δ را برآوردگر مجاز گویند.
به بیانی دیگر هرگاه هیچ برآوردگری موجود نباشد که از δ برتر باشد، گوییم δ یک برآوردگر مجاز است و اگر δ برآوردگری مجاز نباشد آن را غیرمجاز می‌خوانیم.

چون مجاز بودن می‌تواند خصوصیت مطلوبی باشد پس تعیین تمام برآوردگرهای مجاز مورد توجه است. البته باید این نکته را مورد توجه قرار داد که مجاز بودن به تنهایی یک معیار ضعیف برای بهینگی است و برآوردگرهای مجاز فراوانی می‌توان یافت. از جمله برآوردگر ثابت $\delta(X) = \theta_0$ که هیچ اطلاعی از X را به کار نمی‌برد، مجاز است. پس تنها بر اساس معیار مجاز بودن نمی‌توانیم این مسئله، که کدام برآوردگر باید انتخاب شود را حل کنیم (زین‌الدینی، ۱۳۸۸).

تعریف آ.۶.۰. برآوردگر مینیماکس

به برآوردگر δ^M که ماکزیمم مخاطره را مینیمم می‌کند، را برآوردگر مینیماکس گویند. یعنی

$$\inf_{\delta} \sup_{\theta} R(\theta, \delta) = \sup_{\theta} R(\theta, \delta^M).$$

به عبارت دیگر برآوردگر را مینیماکس گویند هرگاه به ازای هر برآوردگر δ داشته باشیم

$$\sup_{\theta} R(\theta, \delta^M) \leq \sup_{\theta} R(\theta, \delta).$$

تعریف آ.۰.۷. توزیع نرمال چندمتغیره

بردار p مولفه‌ای X دارای توزیع نرمال p - متغیره با میانگین $\mu \in R^p$ و ماتریس کواریانس Σ است و با $X \sim N_p(\mu, \Sigma)$ نمایش می‌دهند اگر و فقط اگر به ازای هر بردار p مولفه‌ای ثابت a، $a^T X$ دارای نرمال یک متغیره باشد.

اگر Σ یک ماتریس معین مثبت باشد، آنگاه X دارای تابع چگالی احتمال به صورت زیر است

$$f(x) = \frac{|\Sigma|^{-۱/۲}}{(۲\pi)^{p/۲}} exp\Big[\frac{-(x-\mu)^T \Sigma^{-۱}(x-\mu)}{۲}\Big],$$

که در آن $|\Sigma|^{۱/۲}$ ریشه دوم دترمینان Σ است. توجه کنید اگر $p = ۱$، آنگاه صورت درجه دوم در توان نمایی، به $(x-\mu)^۲/\sigma^۲$ تبدیل می‌شود و $X \sim N(\mu, \sigma^۲)$.

تعریف آ.۰.۸. استاندارد کردن اگر نمونه iام مشاهدات را به ازای $i = ۱, ۲, ..., n$ با $x_i = (x_{i۱}, ..., x_{ij}, ..., x_{id})^T$ نمایش دهیم در این صورت استاندارد شده‌ی x_{ij} عبارتست از

$$z_{ij} = \frac{x_{ij} - c_j}{sj}$$

که در آن

$$c_j = \frac{۱}{n} \sum_{i=۱}^{n} x_{ij} \qquad j = ۱, ..., d$$

مقدار میانگین و

$$s_j = \sqrt{\frac{۱}{n-۱} \sum_{i=۱}^{n} (x_{ij} - c_j)^۲}$$

انحراف معیار متغیر jام می‌باشد. z_{ij}ها دارای میانگین صفر و انحراف معیار یک می‌باشند.

پیوست ب
برنامه‌های کامپیوتری

در این پیوست دستورهای مورد نیاز برای خروجی‌ها و نمودارهای ارائه‌شده در فصل‌های ۲ و ۳ با استفاده از نرم‌افزار R آمده است.

دستور زیر شامل تعریف شاخص رند می‌باشد که برای تمام دستورها لازم به اجراست.

```
rand<-function(l,h){
a<-dist(l)
b<-dist(h)
n11<-sum((a==0)&(b==0))
n00<-sum((a!=0)&(b!=0))
n01<-sum((a!=0)&(b==0))
n10<-sum((a==0)&(b!=0))
r<-(n00+n11)/length(a)
}
```

دستور ۰۱. برای محاسبه‌ی مقادیر شاخص رند و رسم نمودار برای خوشه‌بندی از دو روش سنتی و انقباضی با دو خوشه با داده‌های ناهمبسته‌ی ۵ بعدی در حالتی که واریانس‌های درون خوشه‌ای مختلف مورد بررسی است از دستورات زیر استفاده می‌کنیم:

```
ordinary<-shrinkage<-numeric()
```

```
sigvec<-seq(0,10,2)

sigvec[1]<-.1

sign<-length(sigvec)

for (c in 1:sign){

rand1<-rand2<-numeric()

n<-25

sig<-sigvec[c]

p<-5

cl<-2

h1<-rep(1:cl,rep(n,cl))

cov<-diag(sig,p)

phat<-sum(diag(cov))/max(eigen(cov)$values)

identity<-matrix(sig,p,p)

for (o in 1:5000){

iter<-0

x1<-mvrnorm(n,rep(0,p),cov)

x2<-mvrnorm(n,rep(2,p),cov)

x<-rbind(x1,x2)

w<-dim(x)[1]

meanx<-apply(x,2,mean)

bool<-T

km<-kmeans(x,cl,algorithm = "MacQueen")

temp<-km$centers

km<-kmeans(x,temp,iter.max=1,algorithm =  "MacQueen")
```

```
l0<-km$cluster

cent0<-km$centers

randi<-0

while ((randi!=1)&(iter<10)){

      matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)

      mat<-cent0-matmean

      prod<-mat%*%solve(cov)%*%t(mat)

      den<-diag(prod)

      sh<- 1-((phat-2)/(den))

      sh[sh<0]<-0

cn<-meanx+t(sh*(cent0-matmean))

      cent0<-t(cn)

      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)

      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))

km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")

l1<-km$cluster

      randi<-rand(l1,l0)

      l0<-l1

      cent0<-km$centers

iter<-iter+1

}

rand1[o]<-rand(l1,h1)

km<-kmeans(x,cl,iter.max=iter,algorithm="MacQueen")

l2<-km$cluster
```

```
rand2[o]<-rand(l2,h1)
}
shrinkage[c]<-mean(rand1)
ordinary[c]<-mean(rand2)
}
shrinkage
ordinary
plot(sigvec,shrinkage,ylim=c(min(ordinary),1),col=gray(0),
pch=17,lwd=1.75,xlab="Sigma",ylab="Rand","o")
lines(sigvec,ordinary,ylim=c(min(ordinary),1),col=gray(0.4),pch=19,"o")
for (i in 1:2){
  legend(5,i,c('shrinkage','ordinary'),
  lwd=1,pch=c(17,19),col=c(gray(0),gray(.4)))
}
```

دستور ۲. برای محاسبه‌ی مقادیر شاخص رند و رسم نمودار برای خوشه‌بندی از دو روش سنتی و انقباضی با دو خوشه با داده‌های همبسته‌ی ۵ بعدی در حالتی که واریانس‌های درون خوشه‌ای مختلف مورد بررسی است از دستورات زیر استفاده می‌کنیم:

```
ordinary<-shrinkage<-numeric()
sigvec<-seq(0,10,2)
sigvec[1]<-.1
sign<-length(sigvec)
for (c in 1:sign){
rand1<-rand2<-numeric()
n<-25
```

```
r<-.25

sig<-sigvec[c]

p<-5

cl<-2

h1<-rep(1:cl,rep(n,cl))

cov<-matrix(r*sig,p,p)

diag(cov)<-sig

phat<-sum(diag(cov))/max(eigen(cov)$values)

identity<-matrix(sig,p,p)

for (o in 1:5000){

iter<-0

x1<-mvrnorm(n,rep(0,p),cov)

x2<-mvrnorm(n,rep(2,p),cov)

x<-rbind(x1,x2)

w<-dim(x)[1]

meanx<-apply(x,2,mean)

bool<-T

km<-kmeans(x,cl,algorithm = "MacQueen")

temp<-km$centers

km<-kmeans(x,temp,iter.max=1,algorithm =  "MacQueen")

l0<-km$cluster

cent0<-km$centers

randi<-0

while ((randi!=1)&(iter<10)){
```

```
      matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)
      mat<-cent0-matmean
      prod<-mat%*%solve(cov)%*%t(mat)
      den<-diag(prod)
      sh<- 1-((phat-2)/(den))
      sh[sh<0]<-0
cn<-meanx+t(sh*(cent0-matmean))
      cent0<-t(cn)
      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)
      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))
km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
l1<-km$cluster
      randi<-rand(l1,l0)
      l0<-l1
      cent0<-km$centers
iter<-iter+1
}
rand1[o]<-rand(l1,h1)
km<-kmeans(x,cl,iter.max=iter,algorithm="MacQueen")
l2<-km$cluster
rand2[o]<-rand(l2,h1)
}
shrinkage[c]<-mean(rand1)
ordinary[c]<-mean(rand2)
```

```
}
shrinkage
ordinary
plot(sigvec,shrinkage,ylim=c(min(ordinary),1),col=gray(0),
pch=17,lwd=1.75,xlab="Sigma",ylab="Rand","o")
lines(sigvec,ordinary,ylim=c(min(ordinary),1),col=gray(0.4),pch=19,"o")
for (i in 1:2){
  legend(5,i,c('shrinkage','ordinary'),lwd=1,
  pch=c(17,19),col=c(gray(0),gray(.4)))
}
```

دستور ۳. برای محاسبه‌ی مقادیر شاخص رند و رسم نمودار برای خوشه‌بندی از دو روش سنتی و انقباضی با دو خوشه با داده‌های ناهمبسته‌ی ۵ بعدی در حالتی که تغییرات بعد مؤثر ماتریس کواریانس مورد بررسی است از دستورات زیر استفاده می‌کنیم:

```
ordinary<-shrinkage<-numeric()
lamvec<-c(20.4,6.67,4.4,2.86,2.22,1.82,1.54,1.33,1.18)
lamn<-length(lamvec)
for (c in 1:lamn){
rand1<-rand2<-numeric()
n<-25
sig<-4
p<-5
cl<-2
h1<-rep(1:cl,rep(n,cl))
cov<-diag(sig,p)
```

```
cov[1,1]<-lamvec[c]*sig

phat<-sum(diag(cov))/max(eigen(cov)$values)

identity<-matrix(sig,p,p)

for (o in 1:5000){

iter<-0

x1<-mvrnorm(n,rep(0,p),cov)

x2<-mvrnorm(n,rep(2,p),cov)

x<-rbind(x1,x2)

w<-dim(x)[1]

meanx<-apply(x,2,mean)

bool<-T

km<-kmeans(x,cl,algorithm = "MacQueen")

temp<-km$centers

km<-kmeans(x,temp,iter.max=1,algorithm =  "MacQueen")

l0<-km$cluster

cent0<-km$centers

randi<-0

while ((randi!=1)&(iter<10)){

      matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)

      mat<-cent0-matmean

      prod<-mat%*%solve(cov)%*%t(mat)

      den<-diag(prod)

      sh<- 1-((phat-2)/(den))

      sh[sh<0]<-0
```

```
cn<-meanx+t(sh*(cent0-matmean))
      cent0<-t(cn)
      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)
      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))
km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
l1<-km$cluster
      randi<-rand(l1,l0)
      l0<-l1
      cent0<-km$centers
iter<-iter+1
}
rand1[o]<-rand(l1,h1)
km<-kmeans(x,cl,iter.max=iter,algorithm="MacQueen")
l2<-km$cluster
rand2[o]<-rand(l2,h1)
}
shrinkage[c]<-mean(rand1)
ordinary[c]<-mean(rand2)
}
shrinkage
ordinary
```

دستور ۴. برای محاسبه‌ی مقادیر شاخص رند و رسم نمودار برای خوشه‌بندی از سه روش سنتی، انقباضی معمولی و جیمز-استاین تعمیم‌یافته با دو خوشه با داده‌های ناهمبسته‌ی ۵ بعدی در حالتی که واریانس‌های درون خوشه‌ای مختلف مورد بررسی است از دستورات زیر استفاده می‌کنیم:

```
ordinary<-shrinkage0<-shrinkage1<-shrinkage10<-

shrinkage100<-shrinkage1000<-shrinkagep<-numeric()

lo<-100

n<-25

p<-5

cl<-2

h1<-rep(1:cl,rep(n,cl))

sigvec<-seq(0,20,2)

sigvec[1]<-.1

sign<-length(sigvec)

for (c in 1:sign){

randb<-rand3<-numeric()

sig<-sigvec[c]

cov<-diag(sig,p)

phat<-sum(diag(cov))/max(eigen(cov)$values)

b<-c(1000,100,10,1,0,phat-2)

lb<-length(b)

identity<-matrix(sig,p,p)

for (o in 1:lo){

x1<-mvrnorm(n,rep(0,p),cov)

x2<-mvrnorm(n,rep(2,p),cov)

x<-rbind(x1,x2)

meanx<-apply(x,2,mean)

matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)
```

```
km<-kmeans(x,cl,iter.max=1,algorithm =  "MacQueen")
l0<-lt<-km$cluster
cent0<-centt<-km$centers
save<-10
for (k in 1:lb){
l0<-lt
cent0<-centt
randi<-0
iter<-0
while ((randi!=1)&(iter<save)){
      mat<-cent0-matmean
      prod<-mat%*%solve(cov)%*%t(mat)
      den<-diag(prod)
      sh<- 1-((phat-2)/(den+b[k]))
      sh[sh<0]<-0
cn<-meanx+t(sh*(cent0-matmean))
      cent0<-t(cn)
      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)
      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))
km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
l1<-km$cluster
      randi<-rand(l1,l0)
      l0<-l1
      cent0<-km$centers
```

```
      iter<-iter+1
      }
save<-iter
randb[k+(o-1)*lb]<-rand(l1,h1)
}
km<-kmeans(x,cl,iter.max=save,algorithm="MacQueen")
l2<-km$cluster
rand3[o]<-rand(l2,h1)
}
shrinkage1000[c]<-mean(randb[seq(1,(lo*lb),lb)])
shrinkage100[c]<-mean(randb[seq(2,(lo*lb),lb)])
shrinkage10[c]<-mean(randb[seq(3,(lo*lb),lb)])
shrinkage1[c]<-mean(randb[seq(4,(lo*lb),lb)])
shrinkage0[c]<-mean(randb[seq(5,(lo*lb),lb)])
shrinkagep[c]<-mean(randb[seq(6,(lo*lb),lb)])
ordinary[c]<-mean(rand3)
}
shrinkage0
shrinkage1
shrinkage10
shrinkage100
shrinkage1000
shrinkagep
ordinary
```

```
plot(sigvec,shrinkage0,ylim=c(min(ordinary),1),col=gray(0),
pch=17,lwd=1.75,xlab="Sigma",ylab="Rand","o")
lines(sigvec,shrinkage1,ylim=c(min(ordinary),1),
col=gray(.75/2),pch=15,"o")
lines(sigvec,ordinary,ylim=c(min(ordinary),1),
col=gray(0.75),pch=19,"o")
for (i in 1:3){
  legend(5,i,c('shrinkage0','shrinkage1','ordinary'),lwd=1,
  pch=c(17,15,19),col=c(gray(0),gray(.75/2),gray(.75)))
}
```

دستور ۵. برای محاسبه‌ی مقادیر شاخص رند و رسم نمودار برای خوشه‌بندی از سه روش سنتی، انقباضی معمولی و بارانچیک با دو خوشه با داده‌های ناهمبسته‌ی ۵ بعدی در حالتی که واریانس‌های درون خوشه‌ای مختلف مورد بررسی است از دستورات زیر استفاده می‌کنیم:

```
ordinary<-james_stein<-baranchit<-numeric()
n<-25
p<-5
cl<-2
h1<-rep(1:cl,rep(n,cl))
sigvec<-seq(0,50,2)
sigvec[1]<-.1
sign<-length(sigvec)
for (c in 1:sign){
rand1<-rand2<-rand3<-numeric()
sig<-sigvec[c]
```

```
cov<-diag(sig,p)

phat<-sum(diag(cov))/max(eigen(cov)$values)

identity<-matrix(sig,p,p)

for (o in 1:20){

x1<-mvrnorm(n,rep(0,p),cov)

x2<-mvrnorm(n,rep(2,p),cov)

x<-rbind(x1,x2)

meanx<-apply(x,2,mean)

matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)

km<-kmeans(x,cl,iter.max=1,algorithm ="MacQueen")

l0<-lt<-km$cluster

cent0<-centt<-km$centers

randi<-0

iter<-0

while ((randi!=1)&(iter<10)){

      mat<-cent0-matmean

      prod<-mat%*%solve(cov)%*%t(mat)

      den<-diag(prod)

      sh<- 1-(p*(phat-2)/(den*(den+p)))

      sh[sh<0]<-0

cn<-meanx+t(sh*(cent0-matmean))

      cent0<-t(cn)

      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)

      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))
```

```
km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
l1<-km$cluster
      randi<-rand(l1,l0)
      l0<-l1
      cent0<-km$centers
iter<-iter+1
}
rand1[o]<-rand(l1,h1)
l0<-lt
cent0<-centt
randi<-0
iters<-iter
iter<-0
while ((randi!=1)&(iter!=iters)){
      mat<-cent0-matmean
      prod<-mat%*%solve(cov)%*%t(mat)
      den<-diag(prod)
      sh<- 1-((phat-2)/(den))
      sh[sh<0]<-0
cn<-meanx+t(sh*(cent0-matmean))
      cent0<-t(cn)
      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)
      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))
km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
```

```
l1<-km$cluster
      randi<-rand(l1,l0)
      l0<-l1
      cent0<-km$centers
iter<-iter+1
}
rand2[o]<-rand(l1,h1)
km<-kmeans(x,cl,iter.max=iters,algorithm="MacQueen")
l2<-km$cluster
rand3[o]<-rand(l2,h1)
}
baranchit[c]<-mean(rand1)
james_stein[c]<-mean(rand2)
ordinary[c]<-mean(rand3)
}
baranchit
james_stein
ordinary
plot(sigvec,baranchit,ylim=c(min(james_stein),1),col=gray(0),
pch=17,lwd=1.75,xlab="Sigma",ylab="Rand","o")
lines(sigvec,james_stein,ylim=c(min(james_stein),1),
col=gray(.75/2),pch=15,"o")
lines(sigvec,ordinary,ylim=c(min(james_stein),1),
col=gray(0.75),pch=19,"o")
```

```
for (i in 1:3){
  legend(3,i,c('baranchik','james_stein','ordinary'),lwd=1,
  pch=c(17,15,19),col=c(gray(0),gray(.75/2),gray(.75)))
}
```

دستور ۶. برای محاسبه‌ی مقادیر شاخص رند و رسم نمودار برای مقایسه‌ي خوشه‌بندی توسط دو برآوردگر جیمز-استاین تعمیم‌یافته و بارانچیک با پنج خوشه با داده‌های ناهمبسته‌ی ۵ بعدی در حالتی که واریانس‌های درون خوشه‌ای مختلف مورد بررسی است از دستورات زیر استفاده می‌کنیم:

```
ordinary<-james_stein<-baranchit<-numeric()
n<-10
p<-5
cl<-5
h1<-rep(1:cl,rep(n,cl))
sigvec<-seq(0,50,2)
sigvec[1]<-.1
sign<-length(sigvec)
for (c in 1:sign){
rand1<-rand2<-rand3<-numeric()
sig<-sigvec[c]
cov<-diag(sig,p)
phat<-sum(diag(cov))/max(eigen(cov)$values)
bopt<-phat-2
identity<-matrix(sig,p,p)
for (o in 1:100){
x1<-mvrnorm(n,rep(0,p),cov)
```

```
x2<-mvrnorm(n,rep(2,p),cov)

x3<-mvrnorm(n,rep(-2,p),cov)

x4<-mvrnorm(n,rep(-4,p),cov)

x5<-mvrnorm(n,rep(4,p),cov)

x<-rbind(x1,x2,x3,x4,x5)

meanx<-apply(x,2,mean)

matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)

km<-kmeans(x,cl,iter.max=1,algorithm =  "MacQueen")

l0<-lt<-km$cluster

cent0<-centt<-km$centers

randi<-0

iter<-0

while ((randi!=1)&(iter<10)){

      srt<-sort(cent0)

      if (any(is.na(cent0[1,]))) cent0[1,]<-srt[1:p]

      if (any(is.na(cent0[2,]))) cent0[2,]<-srt[1:p]

      if (any(is.na(cent0[3,]))) cent0[3,]<-srt[1:p]

      if (any(is.na(cent0[4,]))) cent0[4,]<-srt[1:p]

      if (any(is.na(cent0[5,]))) cent0[5,]<-srt[1:p]

      mat<-cent0-matmean

      prod<-mat%*%solve(cov)%*%t(mat)

      den<-diag(prod)

      sh<- 1-(p*(phat-2)/(den*(den+p)))

      sh[sh<0]<-0
```

```
cn<-meanx+t(sh*(cent0-matmean))
      cent0<-t(cn)
      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)
      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))
km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
l1<-km$cluster
      randi<-rand(l1,l0)
      l0<-l1
      cent0<-km$centers
      iter<-iter+1
}
rand1[o]<-rand(l1,h1)
l0<-lt
cent0<-centt
randi<-0
iters<-iter
iter<-0
while ((randi!=1)&(iter!=iters)){
      srt<-sort(cent0)
      if (any(is.na(cent0[1,]))) cent0[1,]<-srt[1:p]
      if (any(is.na(cent0[2,]))) cent0[2,]<-srt[1:p]
      if (any(is.na(cent0[3,]))) cent0[3,]<-srt[1:p]
      if (any(is.na(cent0[4,]))) cent0[4,]<-srt[1:p]
      if (any(is.na(cent0[5,]))) cent0[5,]<-srt[1:p]
```

```
      mat<-cent0-matmean

      prod<-mat%*%solve(cov)%*%t(mat)

      den<-diag(prod)

      sh<- 1-((phat-2)/(den+bopt))

      sh[sh<0]<-0

cn<-meanx+t(sh*(cent0-matmean))

      cent0<-t(cn)

      if (any(dist(cent0)==0)) cent0<-t(apply(cent0,1,function(x)

      (x+mvrnorm(1,rep(0,p),abs(jitter(0,amount=0.001))*identity))))

km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")

l1<-km$cluster

      randi<-rand(l1,l0)

      l0<-l1

      cent0<-km$centers

      iter<-iter+1

}

rand2[o]<-rand(l1,h1)

km<-kmeans(x,cl,iter.max=iters,algorithm="MacQueen")

l2<-km$cluster

rand3[o]<-rand(l2,h1)

}

baranchit[c]<-mean(rand1)

james_stein[c]<-mean(rand2)

ordinary[c]<-mean(rand3)
```

```
}
baranchit
james_stein
ordinary
plot(sigvec,baranchit,ylim=c(min(ordinary),1),col=gray(0),
pch=17,lwd=1.75,xlab="Sigma",ylab="Rand","o")
lines(sigvec,james_stein,ylim=c(min(ordinary),1),
col=gray(.75/2),pch=15,"o")
lines(sigvec,ordinary,ylim=c(min(ordinary),1),
col=gray(0.75),pch=19,"o")
for (i in 1:3){
  legend(5,i,c('baranchik','james_stein','ordinary'),lwd=1,
  pch=c(17,15,19),col=c(gray(0),gray(.75/2),gray(.75)))
}
```

دستور ۷. برای محاسبه‌ی مقادیر شاخص رند برای مقایسه‌ي خوشه‌بندی از دو روش K-میانگین سنتی و جیمز-استاین معمولی برای داده‌های ژن مخمر از دستورات زیر استفاده می‌کنیم:

```
x<-gene
cl<-5
cov<-cov(x)
p<-dim(x)[2]
phat<-sum(diag(cov))/max(eigen(cov)$values)
bopt<-phat-2
w<-dim(x)[1]
h1<-c(rep(1,13),rep(2,39),rep(3,8),rep(4,7),rep(5,11))
```

```
meanx<-apply(x,2,mean)

matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)

km<-kmeans(x,cl,iter.max=1,algorithm =  "MacQueen")

l0<-lt<-km$cluster

cent0<-centt<-km$centers

randi<-0

iter<-0

while ((randi!=1)&(iter<10)){

      mat<-cent0-matmean

      prod<-mat%*%solve(cov)%*%t(mat)

      den<-diag(prod)

      sh<- 1-((phat-2)/(den))

      sh[sh<0]<-0

cn<-meanx+t(sh*(cent0-matmean))

      cent0<-t(cn)

km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")

l1<-km$cluster

      randi<-rand(l1,l0)

      l0<-l1

      cent0<-km$centers

iter<-iter+1

}

shrinkage<-rand(l1,h1)

l0<-lt
```

```
cent0<-centt
randi<-0
iters<-iter
iter<-0
while ((randi!=1)&(iter!=iters)){
  mat<-cent0-matmean
  prod<-mat%*%solve(cov)%*%t(mat)
  den<-diag(prod)
  sh<- 1-(p*(phat-2)/(den*(den+p)))
  sh[sh<0]<-0
  cn<-meanx+t(sh*(cent0-matmean))
  cent0<-t(cn)
  km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
  l1<-km$cluster
  randi<-rand(l1,l0)
  l0<-l1
  cent0<-km$centers
  iter<-iter+1
}
ordinary<-rand(l2,h1)
shrinkage
ordinary
```

دستور ۸. برای محاسبه‌ی مقادیر شاخص رند برای مقایسه‌ي خوشه‌بندی از سه K-میانگین سنتی، روش جیمز-استاین تعمیم‌یافته و بارانچیک برای داده‌های سفال رومانو-بریتیش از دستورات زیر استفاده

می‌کنیم:

```
x<-pottery[,1:9]

cl<-5

p<-dim(x)[2]

cov<-cov(x)

phat<-sum(diag(cov))/max(eigen(cov)$values)

bopt<-phat-2

w<-dim(x)[1]

h1<-pottery[,10]

meanx<-apply(x,2,mean)

matmean<-matrix(rep(meanx,cl),nrow=cl,byrow=T)

km<-kmeans(x,cl,iter.max=1,algorithm =  "MacQueen")

l0<-lt<-km$cluster

cent0<-centt<-km$centers

randi<-0

iter<-0

while ((randi!=1)&(iter<10)){

      mat<-cent0-matmean

      prod<-mat%*%solve(cov)%*%t(mat)

      den<-diag(prod)

      sh<- 1-((phat-2)/(den))

      sh[sh<0]<-0

cn<-meanx+t(sh*(cent0-matmean))

      cent0<-t(cn)
```

```
km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
l1<-km$cluster
      randi<-rand(l1,l0)
      l0<-l1
      cent0<-km$centers
iter<-iter+1
}
baranchik<-rand(l1,h1)
l0<-lt
cent0<-centt
randi<-0
iter<-0
while ((randi!=1)&(iter!=iters)){
  mat<-cent0-matmean
  prod<-mat%*%solve(cov)%*%t(mat)
  den<-diag(prod)
  sh<- 1-((phat-2)/(den+bopt))
  sh[sh<0]<-0
  cn<-meanx+t(sh*(cent0-matmean))
  cent0<-t(cn)
  km<-kmeans(x,cent0,iter.max=1,algorithm ="MacQueen")
  l1<-km$cluster
  randi<-rand(l1,l0)
  l0<-l1
```

```
  cent0<-km$centers
  iter<-iter+1
}
taamim<-rand(l1,h1)
km<-kmeans(x,cl,iter.max=iters,algorithm="MacQueen")
l2<-km$cluster
ordinary<-rand(l2,h1)
baranchik
taamim
ordinary
```

مراجع

[۱] حسن‌زاده بشتیان م، (۱۳۸۵)، پایان‌نامه کارشناسی‌ارشد، "خوشه‌بندی به روش‌های نقطه ثابت، سلسله‌مراتبی و غیرسلسله‌مراتبی "،دانشکده علوم ریاضی، دانشگاه فردوسی مشهد.

[۲] داورزنی غ، (۱۳۸۸)، پایان‌نامه کارشناسی‌ارشد، "الگوریتم‌های تحلیل خوشه‌ای"، دانشکده ریاضی، آمار و کامپیوتر، دانشگاه تهران.

[۳] زین‌الدینی ش، (۱۳۸۸)، پایان‌نامه کارشناسی‌ارشد، "مروری بر برآورد بیزی مینیماکس میانگین در توزیع نرمال چندمتغیره آمیخته در مقیاس"، دانشکده ریاضی، آمار و کامپیوتر، دانشگاه تهران.

[۴] فرشادفر، ع. (۱۳۸۹) "اصول و روشهای آماری چندمتغیره"، کرمانشاه: دانشگاه رازی

[۵] فرهادی ز، (۱۳۹۳)، پایان‌نامه کارشناسی‌ارشد، "خوشه‌بندی توسط روش رده‌بندی جنگل‌های تصادفی"، دانشکده ریاضی، دانشگاه شاهرود.

[۶] نوروزی‌راد م، (۱۳۹۰)، پایان‌نامه کارشناسی‌ارشد، "مباحثی در برآورد تابع زیان"، دانشکده علوم پایه، دانشگاه آزاد اسلامی واحد مشهد.

[7] Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97, 10101-10106.

[8] Baranchik, A.J., 1964. Multiple regression and estimation of the mean of a multivariate normal distribution. *Technical Report* 51, Department of Statistics. Stanford University, Stanford, CA.

[9] Bock, M.E., 1975. Minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* 3, 209-218.

[10] Brusco, M.J., Cradit, J.D., 2001. A variable-selection heuristic for K-means clustering. *Psychometrika* 66, 249-270.

[11] Cuesta-Albertos, J.A., Gordaliza, A., Matrán, C., 1997. Trimmed K-means: an attempt to robustify quantizers. *Ann. Statist.* 25, 553-576.

[12] Cuesta-Albertos, J.A., Matrán, C., Mayo-Iscar, A., 2008. Robust estimation in the normal mixture model based on robust clustering. *J. Royal. Statist. Soc. Ser. B Statist. Methodol* 70, 779-802.

[13] DeSarbo, W.S., Carroll, J.D., Clark, L.A., Green, P.E., 1984. Synthesized clustering: a method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika* 49, 187-215.

[14] De Jager, C., Nieuwenhuijzen, H., Van Der Hucht, K. A., 1988. Mass loss rates in the Hertzsprung-Russell diagram. *Astronomy and Astrophysics Supplement Series*, 72, 259-289.

[15] Everitt B., Hothorn T., 2011. *An Introduction to Applied Multivariate Analysis with R*, Springer, New York.

[16] Forgy, E.W., 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics* 21, 768–769.

[17] Gan, G., Ma, C., Wu, J., 2007. *Data Clustering: Theory, Algorithms, and Applications.* SIAM-ASA, Alexandria, VA.

[18] Gao, J. and Hitchcock, D.B., 2010. James-Stein shrinkage to improve K-means cluster analysis, *Comp. Statist. Data Anal.* 54, 2113-2127.

[19] García-Escudero, L.A., Gordaliza, A., San Martín, R., Van Aelst, S., Zamar, R., 2009. Robust linear clustering. *J. Royal. Statist. Soc. Ser. B Statist. Methodol* 71, 301-318.

[20] Hartigan, J.A., Wong, M.A., 1979. A k-means clustering algorithm. *J. Royal. Statist. Soc. Ser. C. Appl. Statist.* 28, 100-108.

[21] Hitchcock, D.B., Booth, J.G., Casella, G., 2007. The effect of pre-smoothing functional data on cluster analysis. *J. Statist. Comp. Sim.* 77, 1043-1055.

[22] Hitchcock, D.B., Chen, Z., 2008. Smoothing dissimilarities to cluster binary data. *Comput. Statist. Data Anal.* 52, 4699-4711.

[23] James, W., Stein, C., 1961. Estimation with quadratic loss. In: *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* 1, 361-379.

[24] Kaufman, L., Rousseeuw, P.J., 1987. Clustering by means of medoids. In: *Dodge, Y. (Ed.), Statistical Data Analysis Based on the L1-Norm.* North-Holland, Amsterdam, pp. 405-416.

[25] Krzanowski, W.J., Hand, D.J., 2009. A simple method for screening variables before clustering of microarray data. *Comput. Statist. Data Anal.* 53, 2747-2753.

[26] Lehmann E. L., Casella G., 1998. *Theory of Point Estimation*, 2nd Ed., Springer, New York.

[27] MacQueen, J., 1967. Some methods for classification and analysis of multivariate observation. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1. University of California Press, Berkeley, pp. 2811-297.

[28] Makarenkov, V., Legendre, P., 2001. Optimal variable weighting for ultrametric and additive trees and K-means partitioning: methods and software. *J. Classification* 18, 245-271.

[29] Maronna, R., Jacovkis, P.M., 1974. Multivariate clustering procedures with variable metrics. *Biometrics* 30, 499-505.

[30] Munsche, H., Whitaker, H. A., 2012. Eighteenth Century Classification of Mental Illness: Linnaeus, de Sauvages, Vogel, and Cullen. *Cognitive and Behavioral Neurology*, 25(4), 224-239.

[31] Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* 66, 846-850.

[32] Richards, J.A., 1999. *An Introduction to James-Stein Estimation.* Available at http://ssg.mit.edu/group/alumni/johnrich/docs/jse.ps.gz in April 2009.

[33] Stein, C., 1956. Inadmisibility of the usual estimator for the mean of multivariate normal distribution. In: *Proceeding of 3rd Berekley Symposiumon Mathematical Statistics and Probility*,1, University of California Press, Brekley, 197-206.

[34] *SAS Institute Inc.* 1999. SAS OnlineDoc, Version 8. SAS Institute Inc. Cary, NC.

[35] Shao J., 2007. *Mathematical Statistics*, 2nd Ed., Springer, New York.

[36] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9, 3273-3297.

[37] Steinley, D., 2006. k-means clustering: a half-century synthesis. *British J. Math. Statist. Psych.* 59, 1-34.

[38] Steinley, D., Brusco, M.J., 2008a. A new variable weighting and selection procedure for K-means cluster analysis. *Multivariate Behav. Res.* 43, 77-108.

[39] Steinley, D., Brusco, M.J., 2008b. Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika* 73, 125-144.

[40] Tan, P., Steinbach, M., Kumar, V., 2005. *Introduction to Data Mining.* Addison-Wesley, Reading, MA.

[41] Tryon, R., 1939. *Cluster analysis.* New York: McGraw Hill.

[42] Tucker, L. R. 1964. *The extension of factor analysis to three-dimensional matrices. In Contributions to Mathematical Psychology.* Holt, Rinehart and Winston, New York, 110-182.

Aabstract

We study a general algorithm to improve the accuracy in cluster analysis that employs the James-Stein shrinkage effect in k-means clustering. We shrink the centroids of clusters toward the overall mean of all data using a James-Stein-type adjustment, and then the James-Stein shrinkage estimators act as the new centroids in the next clustering iteration until convergence. We compare the shrinkage results to the traditional k-means method. A Monte Carlo simulation shows that the magnitude of the improvement depends on the within-cluster variance and especially on the effective dimension of the covariance matrix. Using the Rand index, we demonstrate that accuracy increases significantly in simulated data and in a real data example.

Keywords : Clustering, K-means, Shrinkage estimator, Baranchik estimator, Rand index

Shahrood University
Faculty Of Mathematical Sciences

Dissertation Submitted in Partial
Fulfillment of The Requirements For The
Degree of Master of Science in
Statistics

Shrinkage estimation in K-means cluster analysis for multivariate normal distribution

Supervisor
Dr. Mohammad Arashi

by
Peyman Barabadi

2015