دانشگاه صنعتی خواجه نصیرالدین طوسی

دانشکده مهندسی کامپیوتر

گزارش فاز دوم پروژه داده‌کاوی پیشرفته

مقایسه و اعتبارسنجی خانواده‌های الگوریتم خوشه‌بندی

مرصاد احمدی — علی حسن زاده

۴۰۳۰۱۲۷۴ — ۴۰۳۰۴۴۶۴

استاد: دکتر پیشگو

تیر ۱۴۰۵ / July ۲۰۲۶

فهرست مطالب ۲
چکیده اجرایی ۳
۱. طراحی آزمایش و قرارداد مقایسه ۳
۲. مطالعه تعیین تعداد خوشه ۴
۲.۱. Elbow، Silhouette و معیارهای داخلی ۴
۲.۲. Gap Statistic ۵
۲.۳. پایداری k ۶
۳. روش‌های Partitioning ۸
۴. روش‌های سلسله‌مراتبی ۱۰
۵. روش‌های چگالی‌محور ۱۱
۶. مدل‌های آمیخته گاوسی ۱۳
۷. Spectral Clustering ۱۴
۸. مقایسه نهایی خانواده‌ها ۱۴
۹. متغیرهای کمکی، توافق و تحلیل خطا ۱۶
۹.۱. معیارهای متغیرهای کمکی ۱۶
۹.۲. توافق الگوریتم‌ها ۱۶
۹.۳. نقاط کم silhouette ۱۷
۱۰. نتیجه و انتقال به فاز سوم ۱۸
پیوست — بازتولیدپذیری ۱۹

Article II. چکیده اجرایی

در این فاز K-Means و MiniBatch K-Means روی ۶۰,۰۲۳ پاسخ‌دهنده، ۵۰ مؤلفه PCA و ده مقدار k اجرا شدند. مقایسه بین‌خانواده‌ای نیز برای جلوگیری از سوگیری نمونه روی یک نمونه ثابت ۳,۰۰۰تایی انجام شد و ۷۸ پیکربندی از پنج خانواده partitioning، سلسله‌مراتبی، چگالی‌محور، مدل‌مبنا و Spectral را پوشش داد.

شواهد انتخاب k هم‌جهت نیستند: Silhouette و جمع رتبه‌ها k=۲، Kneedle مقدار ۶، Gap ۱-SE مقدار ۷ و بیشترین پایداری bootstrap مقدار ۳ را ترجیح می‌دهند. بهترین silhouette معتبر خانواده‌ها متعلق به GMM-tied با k=۲ است، در حالی که BIC مدل GMM-full با k=۳ را برمی‌گزیند. روش‌های چگالی‌محور در grid مستندشده راه‌حل چندخوشه‌ای معتبر تولید نکردند. این اختلاف، ساختار هم‌پوشان و چندمقیاسی داده را نشان می‌دهد و مبنای استفاده از consensus در فاز سوم است.

جمع‌بندی فاز MiniBatch K-Means با k=۲ مرجع تمام‌داده‌ای و بازتولیدپذیر است، اما برچسب مرجع داده محسوب نمی‌شود. فاز سوم باید چند نامزد معتبر را در ماتریس co-association تجمیع کند و عدم‌قطعیت انتخاب k را در تفسیر نگه دارد.

شاهد	نتیجه
Silhouette / جمع رتبه	k=۲
Elbow / Kneedle	k=۶
Gap ۱-SE	k=۷
Bootstrap stability	k=۳
بهترین family silhouette	GMM-tied، k=۲
بهترین GMM BIC	GMM-full، k=۳

Article III. ۱. طراحی آزمایش و قرارداد مقایسه

تمام الگوریتم‌ها روی X_pca حاصل از RobustScaler اجرا شدند. برای K-Means و MiniBatch از کل cohort استفاده شد؛ معیارهای پرهزینه روی نمونه ثابت ۵۰۰۰تایی محاسبه شدند. مقایسه خانواده‌ها روی نمونه مشترک ۳۰۰۰تایی با seed=۲۴۲ انجام شد تا تفاوت امتیازها ناشی از تغییر مشاهده‌ها نباشد. برای هر پیکربندی runtime، تعداد خوشه واقعی، سهم نویز، کمترین و بیشترین سهم خوشه و معیارهای داخلی ثبت شد.

خانواده	الگوریتم‌ها	فرض هندسی
Partitioning	K-Means، MiniBatch	خوشه‌های نسبتا کروی و centroidمحور

خانواده	الگوریتم‌ها	فرض هندسی
Hierarchical	single، complete، average، Ward	ساختار سلسله‌مراتبی و linkage
Density	DBSCAN، HDBSCAN، OPTICS	چگالی متغیر، نویز و شکل نامنظم
Model-based	GMM با ۴ covariance	توزیع گاوسی و عضویت نرم
Graph-based	Spectral nearest-neighbors	ساختار غیرکوژ روی گراف شباهت

قاعده اعتبار پیکربندی فقط زمانی نامزد انتخاب است که بیش از یک خوشه تولید کند و هر خوشه دست‌کم ۲٪ نمونه مشترک را داشته باشد. این قاعده تقسیم single-linkage با ۲۹۹۹ عضو در برابر یک عضو را کنار گذاشت، هرچند silhouette ظاهری آن بسیار بالا بود.

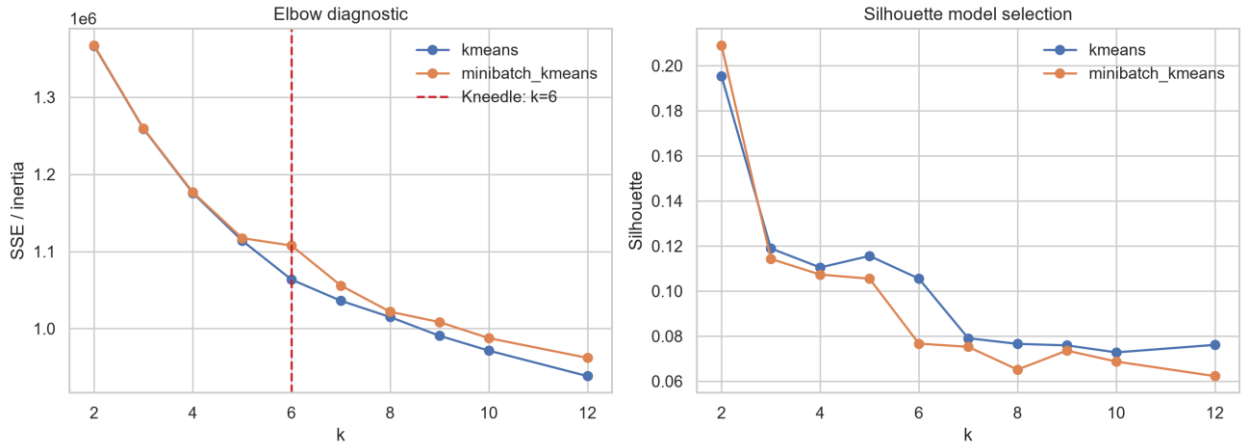
Article IV. ۲. مطالعه تعیین تعداد خوشه

Section ۴.۰۱ ۲.۱. Elbow، Silhouette و معیارهای داخلی

Inertia با افزایش k به‌طور یکنواخت کاهش می‌یابد و به‌تنهایی انتخاب قطعی نمی‌دهد. Kneedle شکست منحنی را در k=۶ تشخیص داد. Silhouette، Davies–Bouldin، Calinski–Harabasz و Dunn تقریبی روی نمونه ارزیابی مشترک محاسبه شدند. بهترین Silhouette و جمع رتبه ترکیبی به k=۲ رسید؛ این نتیجه یک تقسیم درشت را ترجیح می‌دهد.

الگوریتم	k	Silhouette	DB	CH	Dunn	زمان
minibatch_kmeans	۲	۰.۲۰۹۰	۲.۲۸۲	۷۲۲.۰	۰.۰۸۸۵	۰.۲۱s
kmeans	۲	۰.۱۹۵۴	۲.۳۵۰	۷۲۳.۳	۰.۱۵۴۱	۱.۴۶s
kmeans	۳	۰.۱۱۸۹	۲.۳۲۸	۵۹۷.۷	۰.۱۰۶۲	۰.۶۰s
kmeans	۵	۰.۱۱۵۶	۲.۱۱۰	۵۰۰.۳	۰.۰۷۸۸	۰.۹۸s
minibatch_kmeans	۳	۰.۱۱۴۳	۲.۳۵۷	۵۹۸.۰	۰.۰۴۸۶	۰.۱۴s
kmeans	۴	۰.۱۱۰۵	۲.۲۷۴	۵۳۴.۸	۰.۱۷۴۲	۰.۸۸s
minibatch_kmeans	۴	۰.۱۰۷۳	۲.۲۸۲	۵۳۵.۳	۰.۰۵۶۴	۰.۲۴s

الگوریتم	k	Silhouette	DB	CH	Dunn	زمان
kmeans	۶	۰.۱۰۵۵	۲.۱۱۱	۴۷۳.۳	۰.۰۵۹۹	۱.۶۴s
minibatch_kmeans	۵	۰.۱۰۵۵	۲.۱۸۲	۴۹۶.۳	۰.۰۷۵۲	۰.۲۶s
kmeans	۷	۰.۰۷۹۱	۲.۳۵۲	۴۲۱.۹	۰.۰۷۳۶	۱.۹۶s



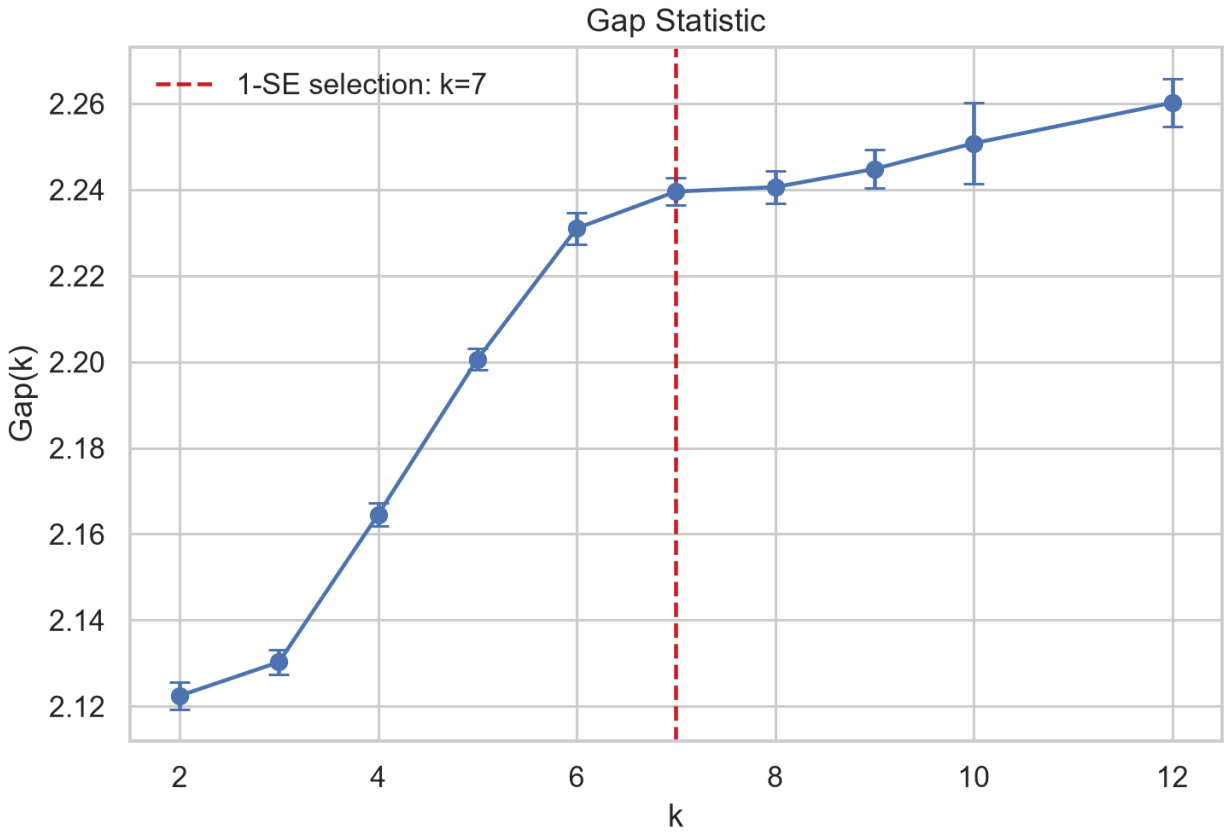
شکل ۱ — شواهد انتخاب k برای K-Means و MiniBatch

Section ۴.۰۲ ۲.۲. Gap Statistic

k	Gap	خطای معیار	شرط SE-۱
۲	۲.۱۲۲۴	۰.۰۰۳۲	False
۳	۲.۱۳۰۳	۰.۰۰۲۹	False
۴	۲.۱۶۴۵	۰.۰۰۲۷	False
۵	۲.۲۰۰۶	۰.۰۰۲۴	False
۶	۲.۲۳۱۰	۰.۰۰۳۶	False
۷	۲.۲۳۹۶	۰.۰۰۳۲	True
۸	۲.۲۴۰۶	۰.۰۰۳۸	True
۹	۲.۲۴۴۸	۰.۰۰۴۵	True

k	Gap	خطای معیار	شرط SE-۱
۱۰	۲.۲۵۰۸	۰.۰۰۹۵	False
۱۲	۲.۲۶۰۲	۰.۰۰۵۶	False

قاعده یک انحراف معیار نخستین k را انتخاب می‌کند که بهبود بعدی نسبت به عدم‌قطعیت برتری معناداری نداشته باشد؛ در این داده k=۷ پیشنهاد شد. اختلاف این نتیجه با Silhouette نشان می‌دهد Gap به زیرساختارهای ریزتر نسبت به جدایش درشت حساس‌تر است.



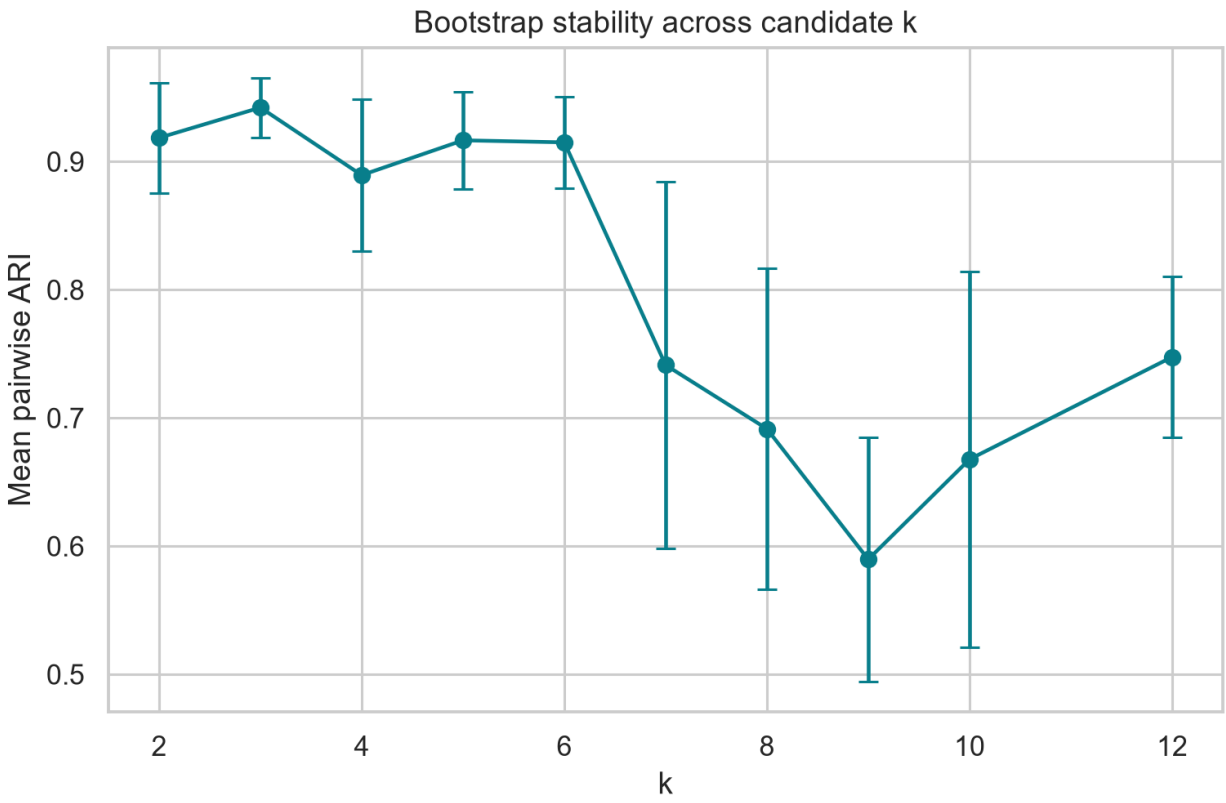
شکل ۲ — Gap Statistic و انتخاب بر اساس قاعده SE-۱

۴.۰۳ Section ۲.۳. پایداری k

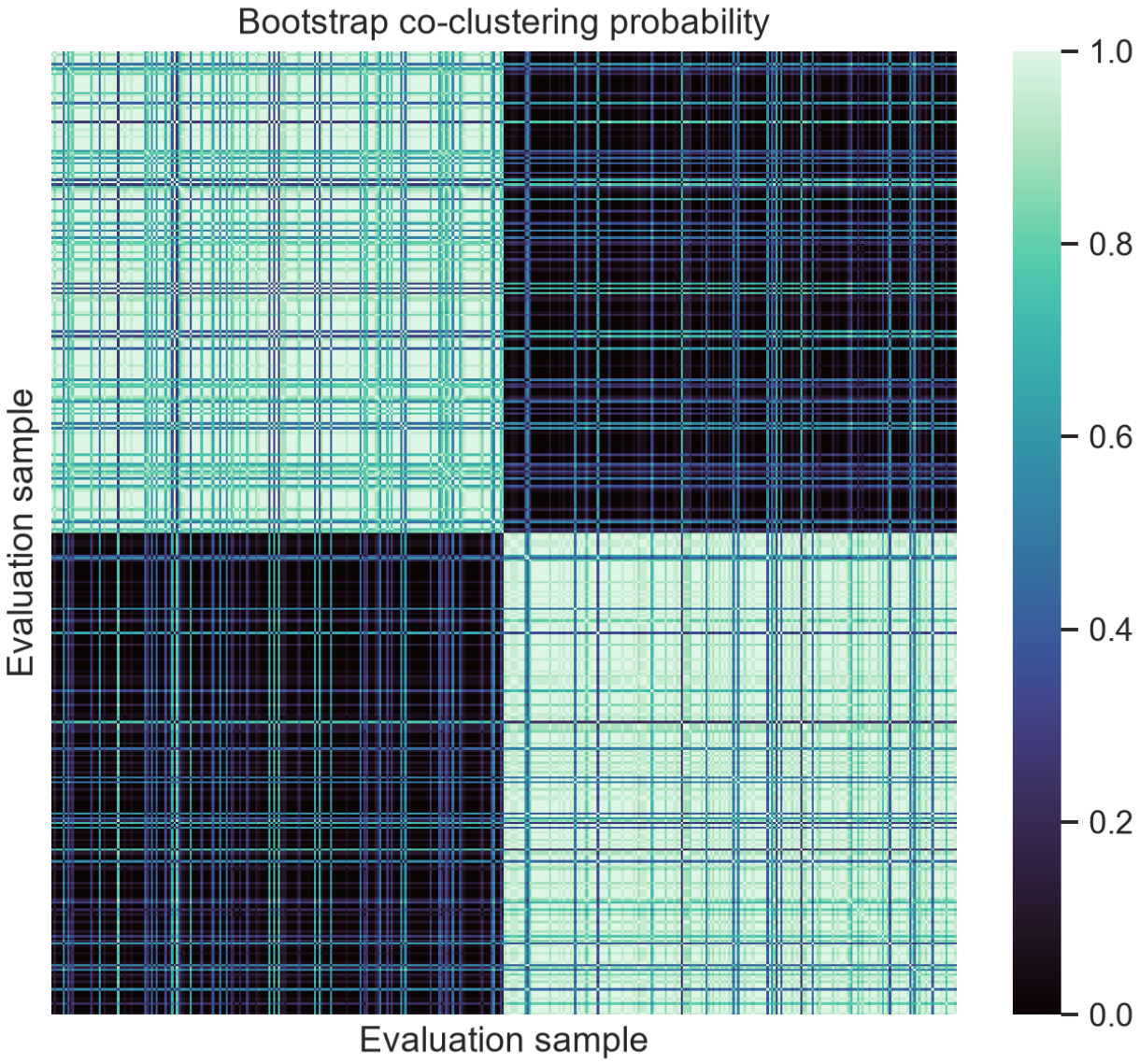
k	تکرار	میانگین ARI	انحراف معیار	کمینه
۲	۸	۰.۹۱۸۶	۰.۰۴۲۸	۰.۸۲۸۹
۳	۸	۰.۹۴۲۲	۰.۰۲۳۲	۰.۸۸۳۵
۴	۸	۰.۸۸۹۴	۰.۰۵۹۳	۰.۷۴۰۹
۵	۸	۰.۹۱۶۶	۰.۰۳۷۸	۰.۸۴۹۷

k	تکرار	میانگین ARI	انحراف معیار	کمینه
۶	۸	۰.۹۱۵۰	۰.۰۳۵۷	۰.۸۳۶۹
۷	۸	۰.۷۴۱۳	۰.۱۴۳۳	۰.۵۷۲۷
۸	۸	۰.۶۹۱۵	۰.۱۲۵۵	۰.۴۶۲۱
۹	۸	۰.۵۸۹۷	۰.۰۹۵۳	۰.۴۱۷۷
۱۰	۸	۰.۶۶۷۸	۰.۱۴۶۴	۰.۳۹۲۶
۱۲	۸	۰.۷۴۷۶	۰.۰۶۲۸	۰.۶۳۰۷

بیشترین میانگین ARI بازنمونه‌گیری برای k=۳ به دست آمد. پایداری بالا به معنی تفسیر دامنه‌ای بهتر نیست، اما نشان می‌دهد تقسیم سه‌خوشه‌ای در برابر حذف تصادفی بخشی از داده مقاوم‌تر است. در انتخاب نهایی این شاهد کنار معیارهای هندسی و تفسیر قرار گرفت.



شکل ۳ — پایداری bootstrap به تفکیک k

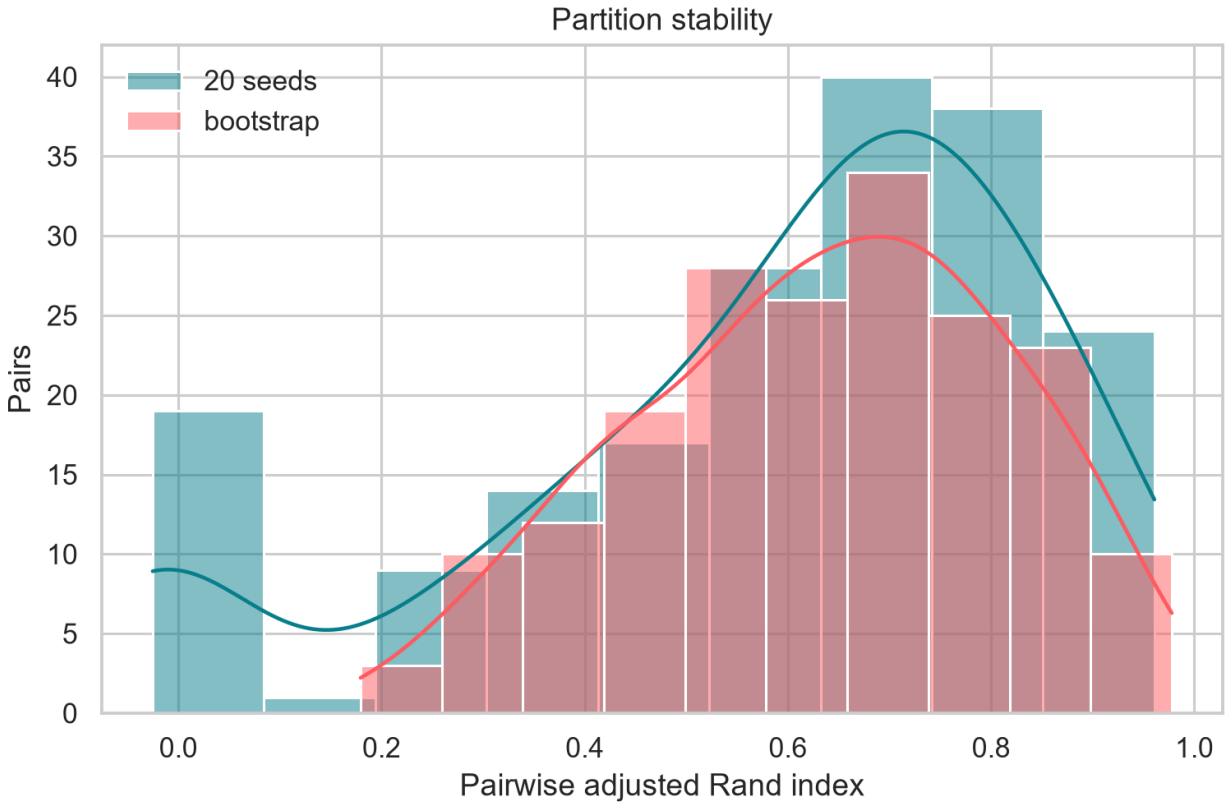


شکل ۴ — احتمال هم‌خوشه‌شدن زوج‌ها در بازنمونه‌گیری bootstrap

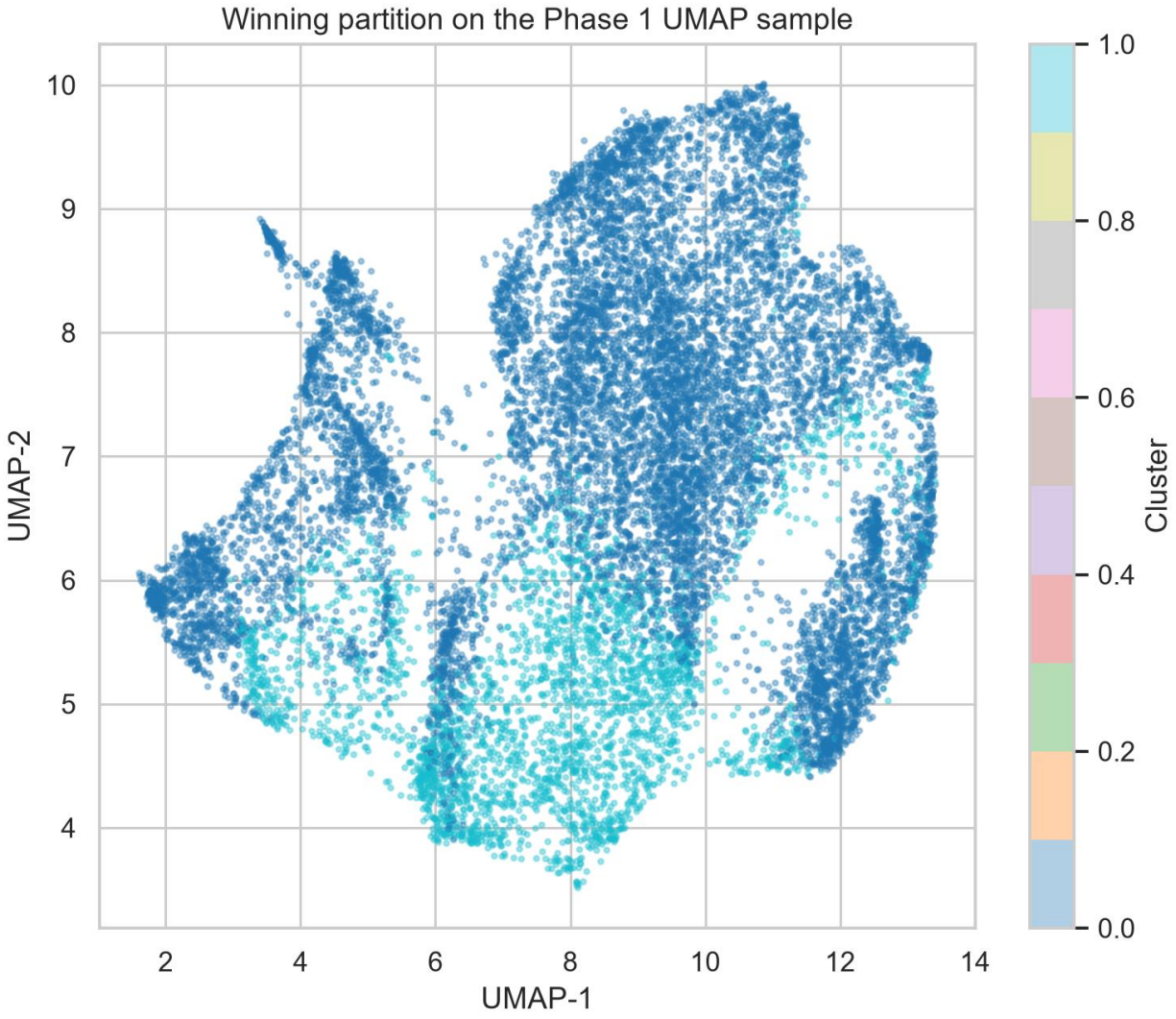
Article V. ۳. روش‌های Partitioning

K-Means با چند شروع مستقل و MiniBatch K-Means با minibatchهای مقیاس‌پذیر اجرا شدند. هر دو الگوریتم برای تمام kها همگرا شدند. مرجع برنده MiniBatch K-Means با k=۲ و silhouette=۰.۲۰۹۰ بود. مدل نهایی روی تمام ۶۰٬۰۲۳ ردیف برازش و هش آن در manifest ثبت شد.

پایداری بیست seed این مرجع میانگین ARI=۰.۵۸۳ و انحراف معیار ۰.۲۶۷ داشت. بیست bootstrap نیز میانگین ۰.۶۳۳ ایجاد کرد. پراکندگی این مقادیر نشان می‌دهد centroidهای درشت قابل بازیابی‌اند، اما نقاط مرزی و زیرساختارها می‌توانند جابه‌جا شوند.



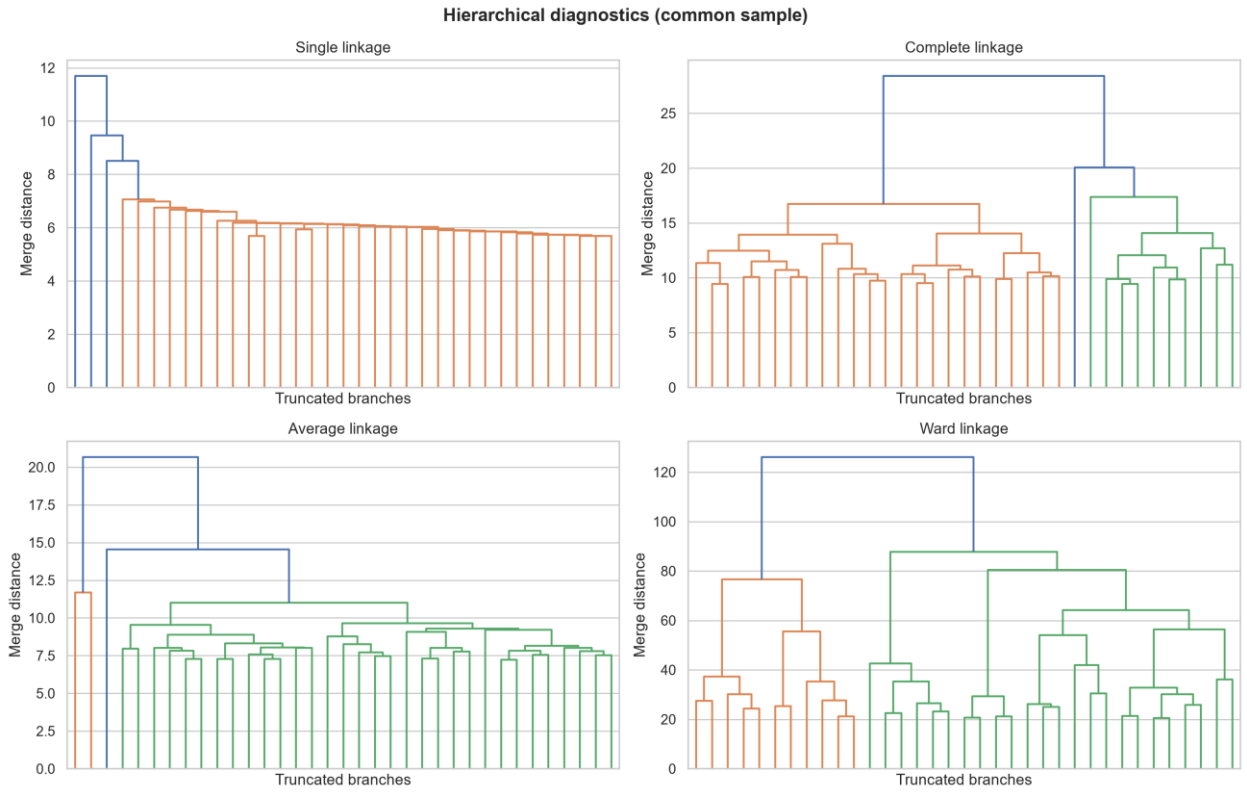
شکل ۵ — توزیع پایداری seed و bootstrap برای مرجع partitioning



شکل ۶ — برچسب مرجع partitioning روی UMAP دوبعدی

Article VI. ۴. روش‌های سلسله‌مراتبی

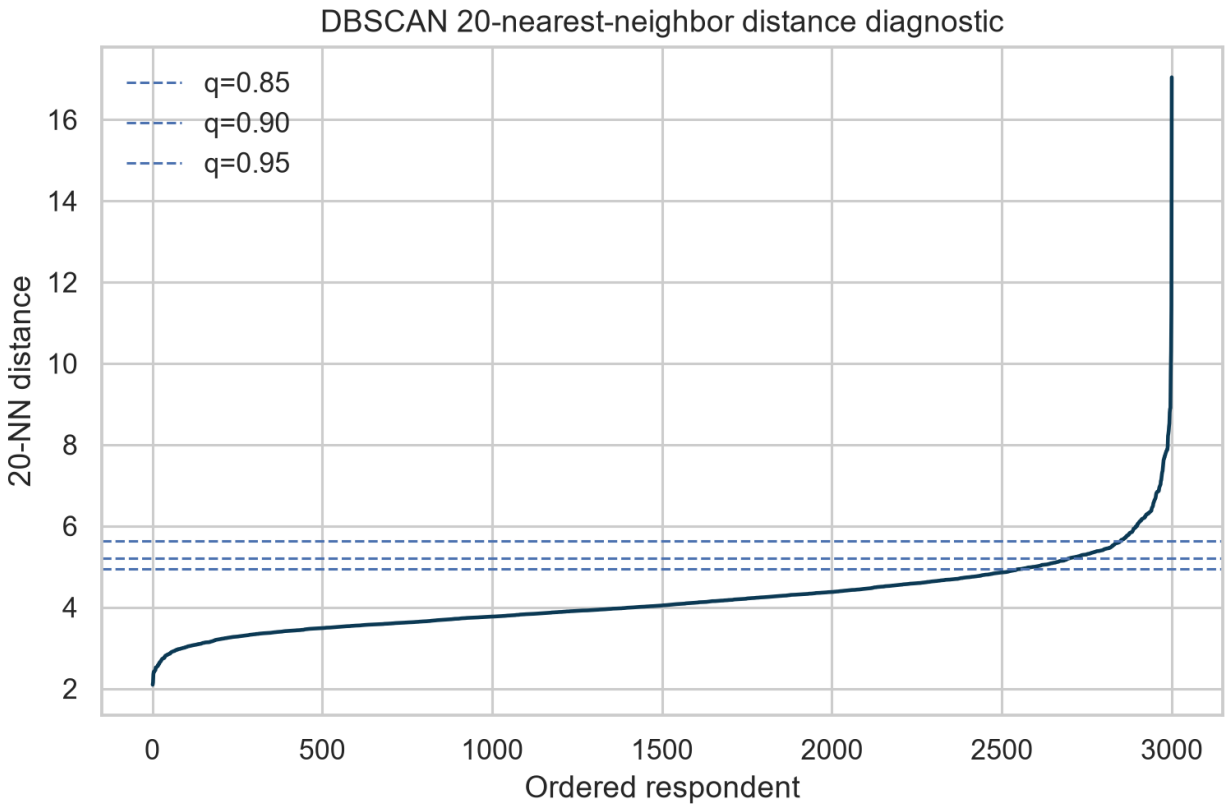
چهار linkage روی نمونه مشترک بررسی شد. single نسبت به chaining آسیب‌پذیر بود و تقسیم بسیار نامتوازن تولید کرد. complete و average خوشه‌های فشرده‌تر ساختند، اما Ward با k=۲ بهترین راه‌حل معتبر این خانواده و silhouette حدود ۰/۲۱۲ را ارائه داد. cophenetic correlation برای سنجش میزان حفظ فاصله‌های زوجی و دو راهبرد برش maxclust و آستانه فاصله ثبت شدند.



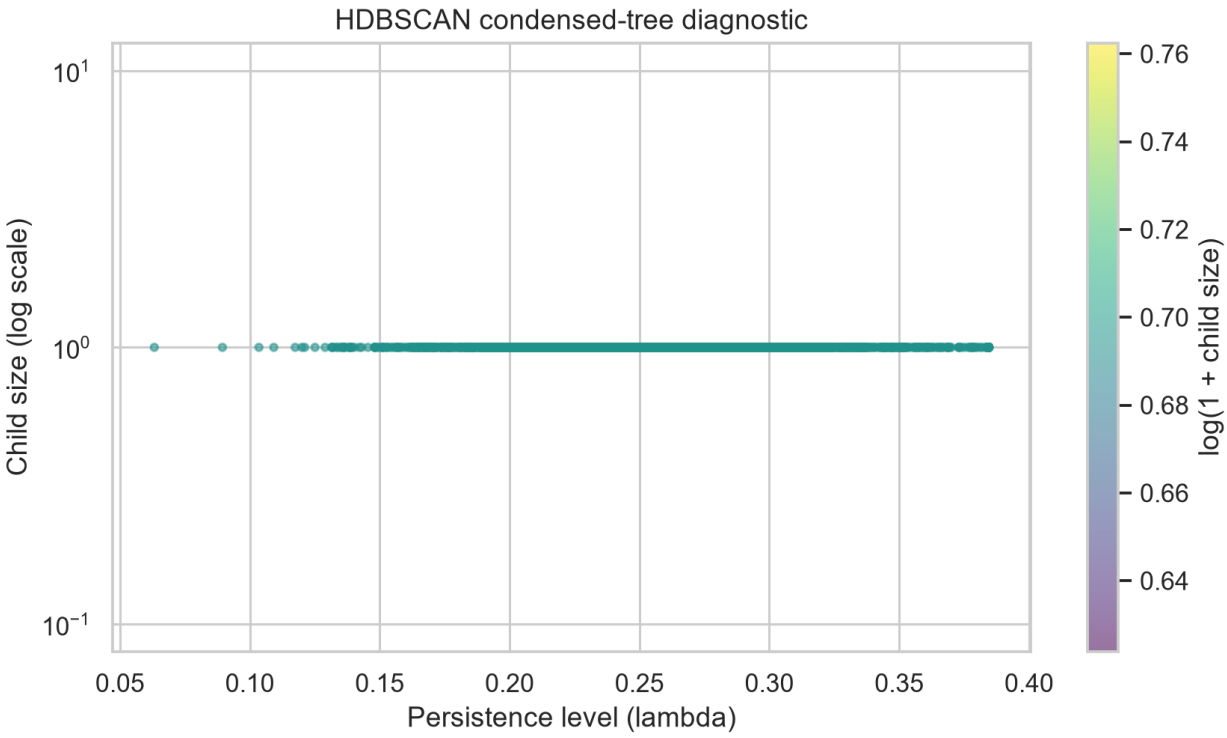
شکل ۷ — دندروگرام‌های چهار linkage سلسله‌مراتبی

Article VII. ۵. روش‌های چگالی‌محور

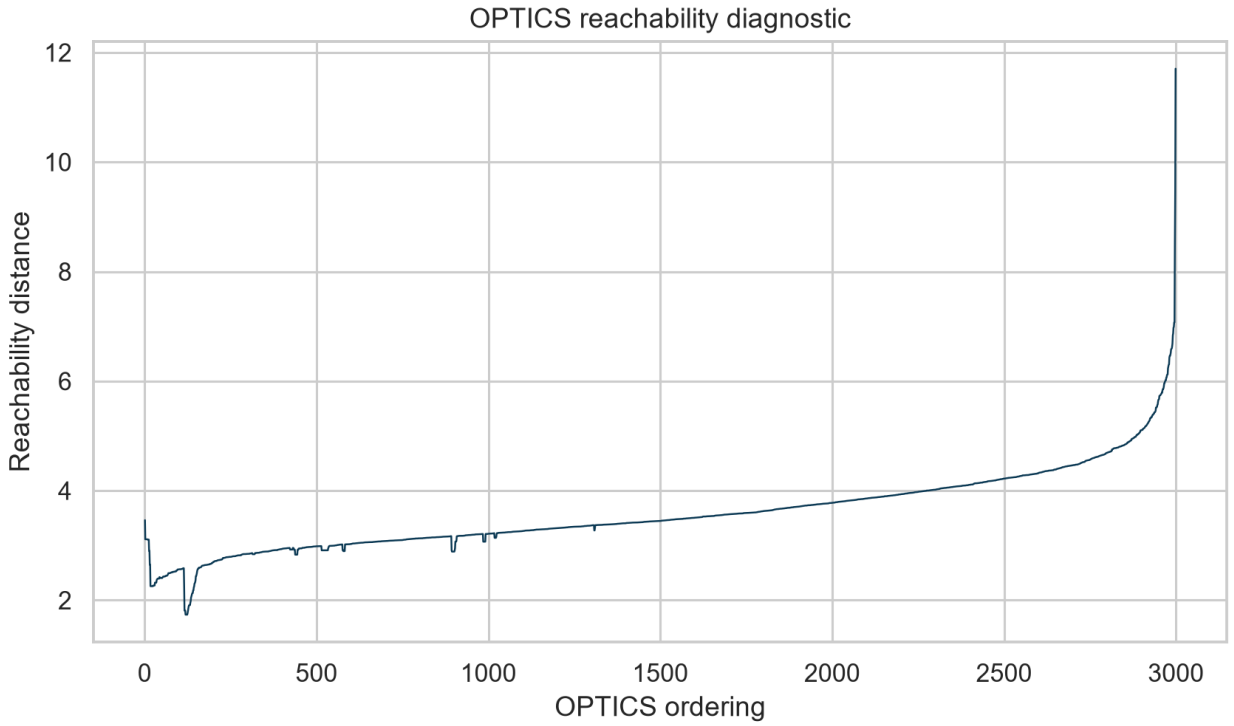
DBSCAN با HDBSCAN، k-distance با condensed tree و OPTICS با reachability curve تنظیم شدند. هدف یافتن شکل‌های نامنظم و جداسازی نویز بود. در grid مستندشده هیچ روش چگالی‌محور راه‌حل چندخوشه‌ای معتبر با حداقل سهم ۲٪ تولید نکرد؛ برخی تنظیم‌ها یک خوشه غالب همراه با نویز و برخی دیگر تقسیم‌های بسیار ریز ساختند. بنابراین در محدوده تنظیمات بررسی‌شده، روش‌های چگالی‌محور برای این داده مناسب نبودند.



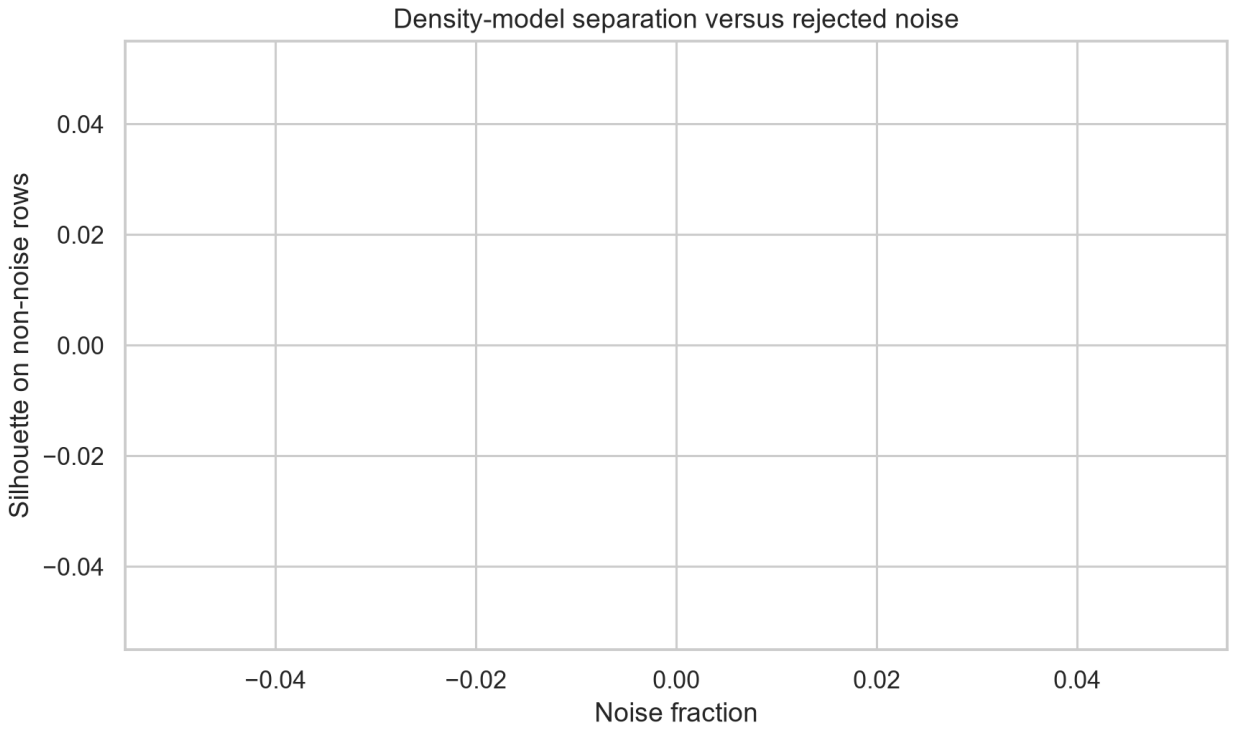
شکل ۸ — منحنی k-distance برای تنظیم روش‌های چگالی‌محور



شکل ۹ — نمودار condensed tree الگوریتم HDBSCAN



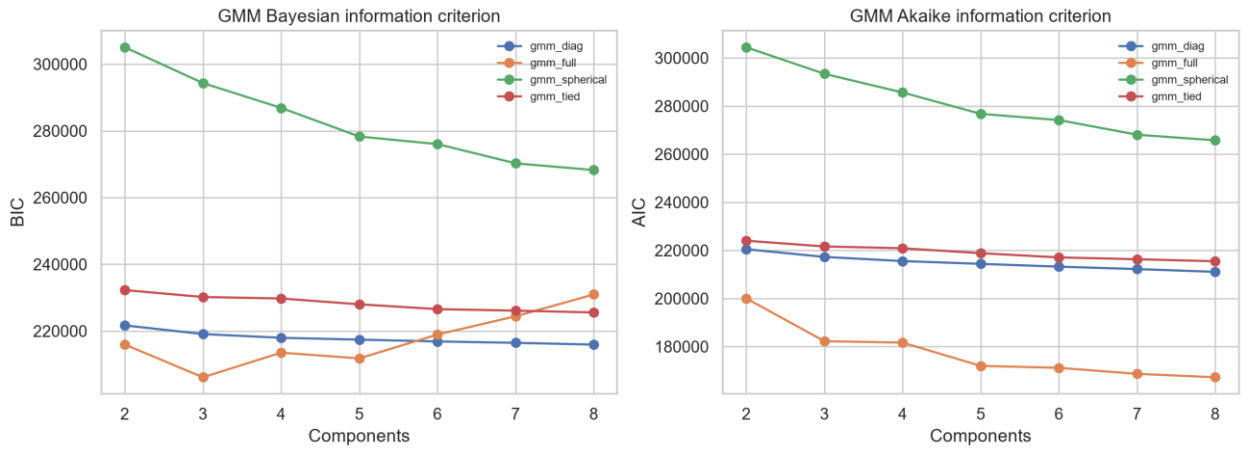
شکل ۱۰ — نمودار reachability الگوریتم OPTICS



شکل ۱۱ — مصالحه جدایش خوشه و سهم نویز

Article VIII. ۶. مدل‌های آمیخته گاوسی

GMM برای covariance‌های tied، diagonal، spherical و full و kهای ۲ تا ۸ اجرا شد. همه مدل‌ها BIC، AIC، همگرایی، lower bound و entropy انتساب نرم را گزارش کردند. GMM-tied با k=۲ بهترین silhouette معتبر خانواده‌ها را داشت، در حالی که کمینه BIC برای GMM-full با k=۳ رخ داد. این تفاوت میان جدایش هندسی و برازش احتمالاتی باید در تفسیر حفظ شود.



شکل ۱۲ — AIC و BIC برای ساختارهای covariance مختلف

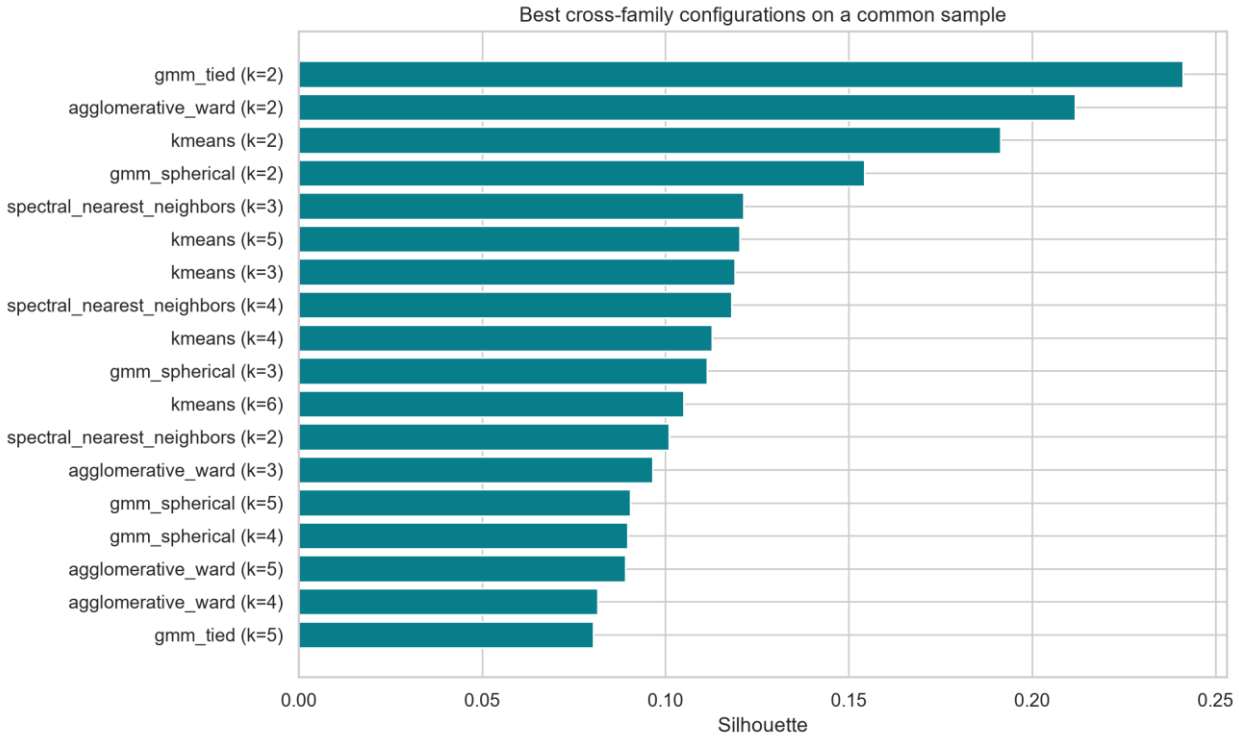
Article IX. ۷. Spectral Clustering

Spectral Clustering با affinity نزدیک‌ترین همسایه روی نمونه مشترک اجرا شد تا ساختارهای غیرکوژ بررسی شوند. بهترین راه‌حل این خانواده k=۳ با silhouette حدود ۰/۱۲۱ بود؛ امتیاز آن از Ward و GMM پایین‌تر است، اما الگوی متفاوتش تنوع مجموعه مدل‌های فاز سوم را بیشتر می‌کند. انتخاب اعضای consensus فقط بر رتبه silhouette متکی نیست و تفاوت فرض‌های هر الگوریتم نیز مهم است.

Article X. ۸. مقایسه نهایی خانواده‌ها

خانواده	الگوریتم	k	Silhouette	DB	CH	کمترین سهم	زمان
model_based	gmm_tied	۲	۰.۲۴۱۱	۱.۹۰۰	۲۹۳.۳	۹.۱٪	۰.۲۲s
hierarchical	agglomerative_ward	۲	۰.۲۱۱۷	۲.۳۶۶	۳۴۳.۴	۱۷.۰٪	۰.۲۲s
partitioning	kmeans	۲	۰.۱۹۱۴	۲.۳۵۸	۴۰۸.۹	۲۲.۹٪	۰.۰۸s
model_based	gmm_spherical	۲	۰.۱۵۴۲	۲.۹۲۱	۳۲۹.۶	۳۴.۱٪	۰.۲۰s
spectral	spectral_nearest_neighbors	۳	۰.۱۲۱۳	۲.۰۰۷	۲۶۱.۹	۶.۵٪	۱.۶۶s
partitioning	kmeans	۵	۰.۱۲۰۳	۲.۰۶۰	۳۰۷.۱	۸.۰٪	۰.۱۳s
partitioning	kmeans	۳	۰.۱۱۸۹	۲.۳۱۳	۳۶۵.۷	۱۹.۸٪	۰.۱۳s
spectral	spectral_nearest_neighbors	۴	۰.۱۱۸۰	۲.۰۶۶	۲۷۲.۸	۶.۴٪	۱.۴۶s
partitioning	kmeans	۴	۰.۱۱۲۶	۲.۲۲۱	۳۳۳.۴	۱۰.۸٪	۰.۱۲s

خانواده	الگوریتم	k	Silhouette	DB	CH	کمترین سهم	زمان
model_based	gmm_spherical	۳	۰.۱۱۱۴	۲.۳۲۷	۳۵۰.۲	۲۰.۸٪	۰.۰۹s
partitioning	kmeans	۶	۰.۱۰۴۹	۲.۰۷۰	۲۸۶.۳	۳.۶٪	۰.۱۵s
spectral	spectral_nearest_neighbors	۲	۰.۱۰۰۹	۱.۸۹۴	۱۵۷.۶	۷.۰٪	۱.۶۷s
hierarchical	agglomerative_ward	۳	۰.۰۹۶۵	۲.۴۹۰	۲۶۹.۹	۱۳.۵٪	۰.۲۲s
model_based	gmm_spherical	۵	۰.۰۹۰۴	۲.۲۲۶	۲۹۱.۹	۶.۸٪	۰.۱۴s
model_based	gmm_spherical	۴	۰.۰۸۹۷	۲.۴۲۳	۳۱۷.۰	۱۳.۴٪	۰.۱۴s



شکل ۱۳ — بهترین پیکربندی‌های معتبر هر خانواده

Article XI. ۹. متغیرهای کمکی، توافق و تحلیل خطا

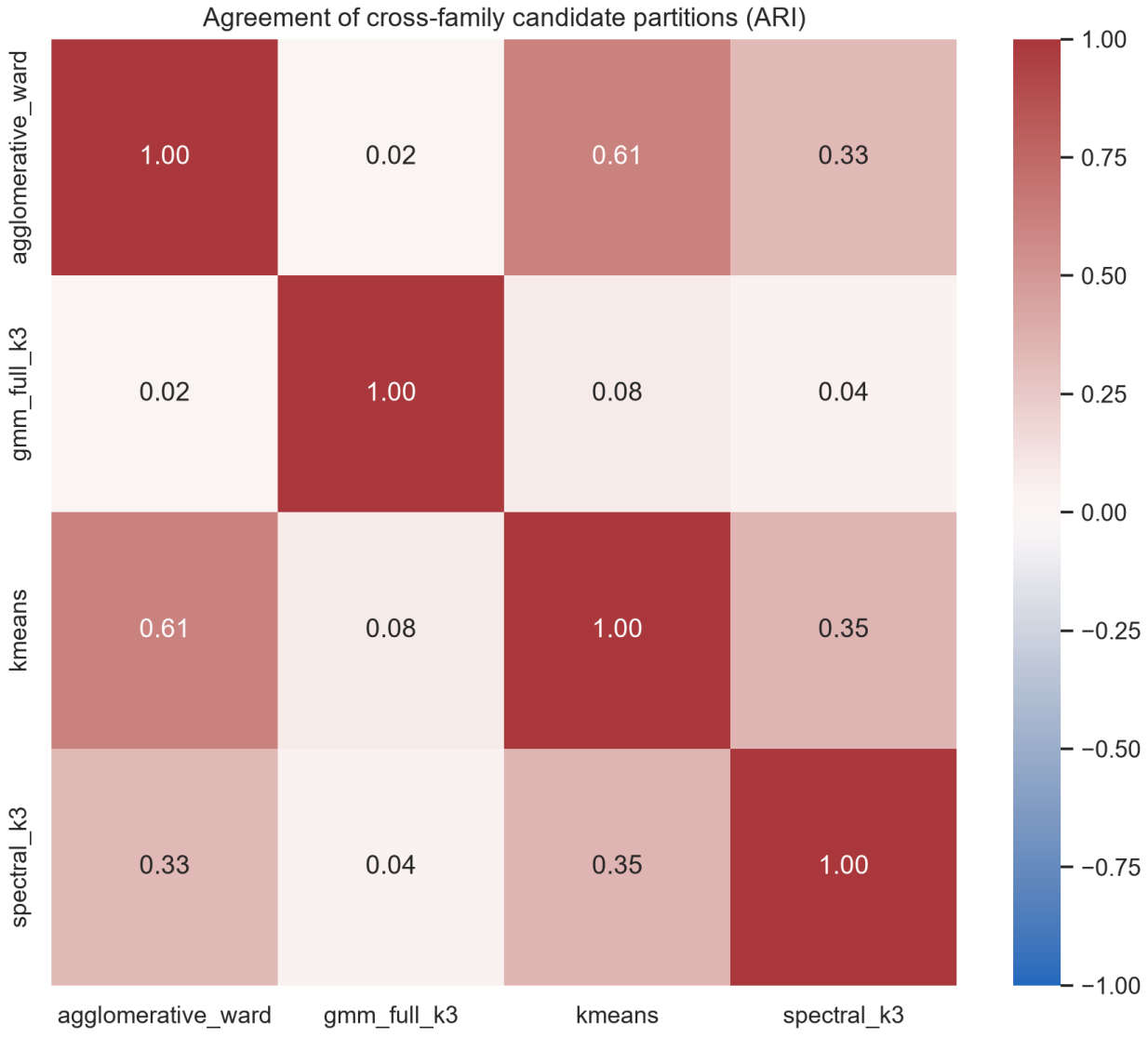
Section ۱۱.۰۱ ۹.۱. معیارهای متغیرهای کمکی

RemoteWork، Employment، EdLevel، DevType و AISelect برچسب مرجع نیستند. NMI فقط نشان می‌دهد خوشه‌ها تا چه اندازه با هر متغیر شناخته‌شده هم‌راستا هستند. مقادیر پایین به‌معنای شکست clustering نیست؛ در واقع نشان می‌دهد تقسیم نهایی صرفا یک نسخه بازکدگذاری‌شده از نقش یا تحصیلات نیست.

Proxy	NMI	دسته	تفسیر
DevType	۰.۰۰۹۳	۳۵	association only; proxy is not ground truth
EdLevel	۰.۰۰۲۲	۹	association only; proxy is not ground truth
Employment	۰.۰۱۲۵	۱۰۳	association only; proxy is not ground truth
RemoteWork	۰.۰۰۱۰	۴	association only; proxy is not ground truth
AISelect	۰.۰۰۳۹	۴	association only; proxy is not ground truth

Section ۱۱.۰۲ ۹.۲. توافق الگوریتم‌ها

نامزد اول	نامزد دوم	ARI
kmeans	agglomerative_ward	۰.۶۱۰۵
kmeans	spectral_k۳	۰.۳۵۱۱
agglomerative_ward	spectral_k۳	۰.۳۲۹۹
kmeans	gmm_full_k۳	۰.۰۸۲۳
gmm_full_k۳	spectral_k۳	۰.۰۴۳۶
agglomerative_ward	gmm_full_k۳	۰.۰۲۱۱



شکل ۱۴ — ماتریس توافق زوجی نامزدهای consensus

Section ۱۱.۰۳ ۹.۳. نقاط کم silhouette

پایین‌ترین silhouetteها به پاسخ‌هایی تعلق دارند که به خوشه خود نزدیک نیستند یا میان centroidها قرار گرفته‌اند. این نقاط برای تحلیل مرز، بازبینی breadth و کنترل وابستگی به تجربه ذخیره شدند. منفی‌بودن silhouette یک فرد به معنی خطای داده نیست؛ نشانه ابهام انتساب در یک ساختار پیوسته است.

ResponseId	خوشه	Silhouette	نقش	تجربه	Breadth
۱۹۰۸۴	۱	-۰.۱۴۰۳	Developer, back-end	۷.۰	۲۵.۰
۱۹۸۶۳	۱	-۰.۱۲۹۰	Other (please specify):	۵.۵	۲۰.۰
۲۵۸۷۰	۱	-۰.۱۲۸۷	Data engineer	۷.۰	۲۷.۰

ResponseId	خوشه	Silhouette	نقش	تجربه	Breadth
۱۵۳۶۳	۱	−۰.۱۲۸۱	Engineer, site reliability	۶.۰	۲۳.۰
۴۳۲۳۴	۱	−۰.۱۲۷۳	Developer, back-end	۷.۰	۲۷.۰
۲۶۱۸۱	۱	−۰.۱۲۵۵	Developer, front-end	۱۰.۵	۲۳.۰
۱۷۲۸۶	۱	−۰.۱۲۵۱	Student	۹.۰	۲۶.۰
۵۴۶۳۰	۱	−۰.۱۲۴۷	Missing	۸.۰	۲۳.۰
۳۶۸۲۳	۱	−۰.۱۲۴۴	Student	۷.۰	۲۳.۰
۷۷۴۲	۱	−۰.۱۲۳۸	Developer, back-end	۳.۰	۲۲.۰
۲۲۶۸۱	۱	−۰.۱۲۳۷	Developer, full-stack	۶.۰	۲۸.۰
۴۱۵۹۷	۱	−۰.۱۲۲۹	System administrator	۵.۵	۲۳.۰

Article XII. ۱۰. نتیجه و انتقال به فاز سوم

یک پاسخ واحد از همه معیارها پشتیبانی نمی‌شود. MiniBatch K-Means با k=۲ مرجع تمام‌داده‌ای است؛ Ward-k۲ جدایش سلسله‌مراتبی معتبر، GMM-full-k۳ برازش احتمالاتی مناسب و Spectral-k۳ ساختار گرافی متفاوت فراهم می‌کنند. این چهار نامزد برای ساخت co-association انتخاب شدند. فاز سوم باید k را در فضای consensus دوباره ارزیابی، برچسب نمونه را به کل داده تعمیم، پروفایل‌ها و نمونه‌های مرزی را تفسیر و حساسیت نتیجه را گزارش کند.

جمع‌بندی انتخاب مدل فاز دوم یک برنده قطعی ندارد و خروجی آن مجموعه‌ای از نامزدهای مناسب همراه با عدم‌قطعیت انتخاب است. این اختلاف‌ها در انتخاب اعضای consensus فاز سوم در نظر گرفته می‌شوند.

```
.\venv\Scripts\python.exe -m scripts.run_phase۲ --config config.yaml
```

تمام seedها، gridها، اندازه نمونه و قاعده حداقل سهم در config.yaml ثبت شده‌اند. برچسب‌ها، linkageها، مدل GMM، مدل partitioning، جدول کامل ۷۸ پیکربندی و شکل‌های تشخیصی در artifacts، data/processed و reports ذخیره شده‌اند.