Негізгі компоненттерді талдау

Айнымалылардың бірге (ковариальді) өзгеруі сирек емес және бір айнымалыдағы вариацияның бір бөлігі екіншісіндегі вариация арқылы дерлік қайталанады. Негізгі компоненттерді талдау (PCA) - сандық айнымалыларды салыстыру әдісін табу әдісі.

Негізгі терминдер:

Негізгі компонент Болжаушы айнымалылардың сызықтық комбинациясы.

Жүктемелер Болжаушыларды құрамдас бөліктерге айналдыруға мүмкіндік беретін салмақтар. Синонимі: салмақ.

Скреплот Компоненттердің салыстырмалы маңыздылығын көрсететін құрамдас дисперсия сызбасы.

PCA-ның идеясы бірнеше сандық болжау айнымалыларын айнымалылардың кішірек жиынына, яғни бастапқы жиынның өлшенген сызықтық комбинацияларына біріктіру болып табылады. Кішірек айнымалылар жиыны (негізгі құрамдас бөліктер) деректердің өлшемділігін азайта отырып, айнымалылардың толық жиынының өзгергіштігінің көп бөлігін «түсіндіреді». Негізгі құрамдастарды қалыптастыру үшін қолданылатын салмақтар бастапқы айнымалылардың жаңа негізгі құрамдастарға қатысты үлестерін көрсетеді.

PCA талдауын Карл Пирсон бастаған. Бақыланбайтын оқыту бойынша бірінші жұмыста Пирсон көптеген тапсырмаларда болжаушы айнымалылардың өзгергіштігі бар екенін мойындады, сондықтан ол осы өзгермелілікті модельдеу әдісін әзірледі. PCA-ны сызықтық дискриминантты талдаудың бақыланбайтын нұсқасы ретінде қарастыруға болады

Қарапайым мысал

X1 және X2 екі айнымалы үшін Z i (i =1 немесе 2) екі негізгі компоненті бар:

$$Z_i = w_{i,1}X_1 + w_{i,2}X_2.$$

Салмақ (wi1, w i2,) құрамдас жүктемелер деп аталады. Олар бастапқы айнымалыларды негізгі құрамдастарға түрлендіреді. Бірінші негізгі компонент, Z1, жалпы вариацияны жақсы түсіндіретін сызықтық комбинация болып табылады. Екінші негізгі компонент, Z2, қалған вариацияны түсіндіреді (бұл да сызықтық комбинация, бірақ ол нашар жуықтау).

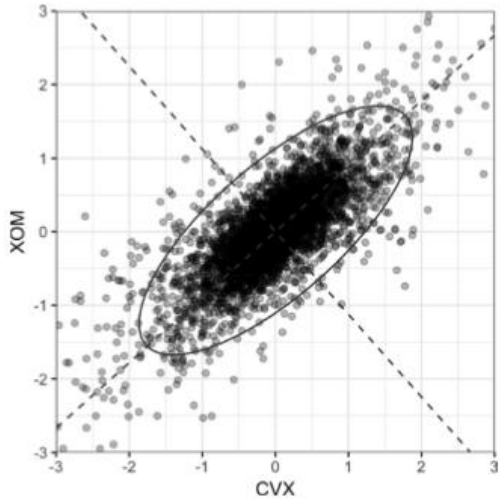
Princomp функциясын пайдаланып R ішіндегі негізгі құрамдастарды есептеуге болады. Келесі код үзіндісі PCA Chevron (CVX) және ExxonMobil (XOM) акцияларының бағасы қайтаруларын орындайды:

```
oil_px <- sp500_px[, c('CVX', 'XOM')]
pca <- princomp(oil_px)
pca$loadings
Loadings:
Comp.1 Comp.2
CVX -0.747 0.665
 XOM -0.665 -0.747
```

CVX және XOM акциялары үшін бірінші негізгі құрамдас үшін салмақтар 0,747 − және 0,665− , ал екінші негізгі компонент үшін салмақтар 0,665 және −0,747. Оны қалай түсіндіруге болады? Бірінші негізгі компонент екі утилита арасындағы корреляцияны көрсететін CVX және XOM мәндерінің орташа мәні болып табылады. Екінші құрамдас CVX және XOM акцияларының бағалары әртүрлі болған кезде өлшенеді. Деректермен бірге негізгі құрамдастарды сызу өте пайдалы:

```
loadings <- pca$loadings
ggplot(data=oil_px, aes(x=CVX, y=XOM)) +
geom_point(alpha=.3) +
stat_ellipse(type='norm', level=.99) +
geom_abline(intercept = 0, slope = loadings[2,1]/loadings[1,1]) +
geom_abline(intercept = 0, slope = loadings[2,2]/loadings[1,2])
```

Нәтиже 7.1.-суретте көрсетілген.



7.1-Сурет. Chevron және ExxonMobil акцияларын қайтару үшін негізгі компоненттер

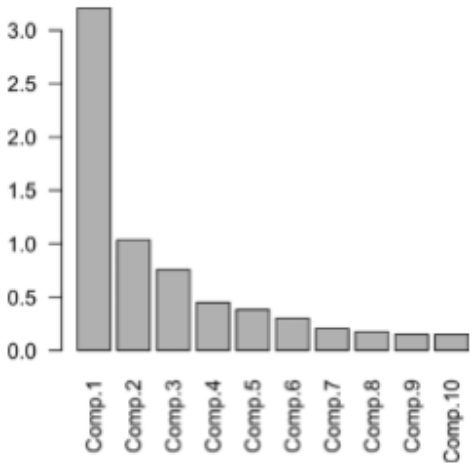
Үзік сызықтар осы екі негізгі құрамдас бөлікті көрсетеді: біріншісі эллипстің ұзын осінің бойында, екіншісі қысқа осьтің бойында. Екі қор кірісіндегі дисперсияның көп бөлігі бірінші негізгі құрамдас бөлікке байланысты екенін көруге болады. Бұл мағынасы бар, өйткені энергия қорларының бағалары топ ретінде қозғалады.

Негізгі құрамдас интерпретация

Негізгі құрамдастардың сипаты көбінесе деректер құрылымы туралы ақпаратты анық етеді. Негізгі құрамдастарды түсінуге көмектесетін бірнеше стандартты визуализация пішіндері бар. Осындай әдістердің бірі негізгі құрамдас бөліктердің салыстырмалы маңыздылығын елестетуге арналған скрипт (Скриплот) болып табылады. Келесі код үзіндісі S&P 500 қор индексіндегі бірнеше үздік компанияларға мысал болып табылады:

```
syms <- c( 'AAPL', 'MSFT', 'CSCO', 'INTC', 'CVX', 'XOM',
 'SLB', 'COP', 'JPM', 'WFC', 'USB', 'AXP', 'WMT', 'TGT', 'HD', 'COST')
top_sp <- sp500_px[row.names(sp500_px)>='2005-01-01', syms]
sp_pca <- princomp(top_sp)
screeplot(sp_pca)
```

7.2-Суреттен көрініп тұрғандай., бірінші негізгі компоненттің дисперсиясы айтарлықтай үлкен, бірақ басқа жоғарғы негізгі құрамдас бөліктер маңызды.



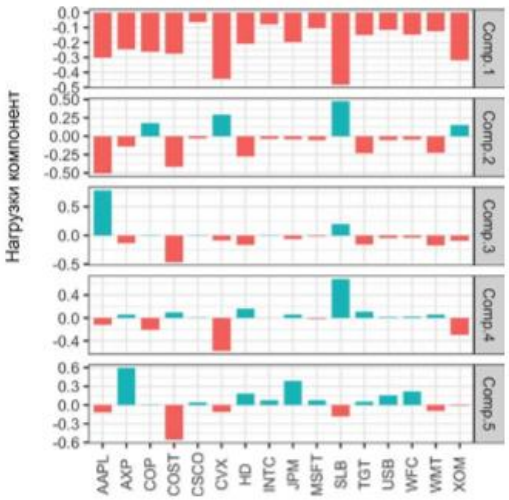
7.2.-сурет S&P 500 жетекші акцияларының PCA үшін скри диаграммасы

Атап айтқанда, жоғарғы негізгі құрамдастардың салмақтарының көрінісі анық болуы мүмкін. Бір жолы tidyr бағдарламалық пакетінен жинау функциясын ggplot-пен бірге пайдалану:

```
library(tidyr) loadings <- sp_pca$loadings[,1:5]
```

```
loadings$Symbol <- row.names(loadings)
loadings <- gather(loadings, "Component", "Weight", -Symbol)
ggplot(loadings, aes(x=Symbol, y=Weight)) +
geom_bar(stat='identity') +
facet_grid(Component ~ ., scales='free_y')
```

Үздік бес құрамдас бөлікке арналған жүктемелер 7.3-суретте көрсетілген. Бірінші негізгі құрамдас бөлікке арналған салмақтар бірдей белгіге ие: бұл барлық бағандардың ортақ мультипликаторы бар деректерге тән (бұл жағдайда қор нарығының жалпы үрдісі). Екінші құрамдас энергия қорларының басқа қорлармен салыстырғандағы баға өзгерістерін қамтиды. Үшінші құрамдас негізінен Apple және CostCo баға көрсеткіштерін салыстырады. Төртінші компонент Schlumberger-тің басқа энергия қорларына қарсы баға әрекетін көрсетеді. Ақырында, бесінші құрамдас бөлікте негізінен қаржы компаниялары басым.



7.3-сурет. Акция бағасының қайтарымының бес негізгі құрамдас бөлігі үшін жүктемелер

Негізгі компоненттерге арналған негізгі идеялар

- *Негізгі құрамдас бөліктер болжау айнымалыларының сызықтық комбинациялары болып табылады (тек сандық деректермен).*
- *Олар артықшылықты азайту кезінде құрамдастардың арасындағы корреляцияны азайту үшін есептеледі.*
- *Құрамдастардың шекті саны әдетте нәтиже айнымалысындағы дисперсияның көп бөлігін түсіндіреді.*
- *Негізгі құрамдастардың бұл шекті жиынын (көптеген) бастапқы болжаушылардың орнына пайдалануға болады, осылайша өлшемді азайтады.*

К-орта негізіндегі кластерлеу

Кластерлеу - бұл әр топтағы жазбалар бір-біріне ұқсас болатындай етіп деректерді әртүрлі топтарға бөлу әдісі. Кластерлеудің мақсаты деректердің маңызды және мағыналы топтарын анықтау болып табылады. Топтарды тікелей пайдалануға, тереңірек талдауға немесе болжамды регрессияға немесе жіктеу үлгісіне мүмкіндік немесе нәтиже ретінде беруге болады. K білдіреді, кластерлеу әдісі ретінде, ең бірінші әзірленді; ол әлі де кеңінен қолданылады және оның танымалдылығы алгоритмнің салыстырмалы қарапайымдылығына және үлкен деректер жиынына масштабтау мүмкіндігіне байланысты.

Негізгі терминдері:

Кластер- Ұқсас жазбалар тобы.

Кластер ортасы- Кластердегі жазбалар үшін айнымалылардың орташа мәндерінің векторы. Синонимдер: центроид, кластер ортасы, масса центрі.

К- Кластер саны.

K-орталары әдісі әрбір жазбаның квадраттық қашықтықтарының қосындысын тағайындалған кластердің орташа мәніне дейін азайту арқылы деректерді K кластерлеріне бөледі. Бұл орта кластер орталығы немесе центроид деп те аталады, квадраттардың кластер ішілік қосындысы немесе кластер ішіндегі SS ретінде көрсетіледі. K кластерлердің өлшемі бірдей болатынына кепілдік бермейді, бірақ ол ең жақсы бөлінген кластерлерді табады.

Қарапайым мысал

n жазбасы және тек екі айнымалысы бар деректер жиынын қарастырудан бастайық, x және y. Біз деректерді K =4 кластерлерге бөлгіміз келеді делік. Бұл әрбір жазба (xi,y i) k- кластеріне жатқызылуы керек дегенді білдіреді. k кластерге n k жазбалардың тағайындалуын ескере отырып, кластердің орталығы (xk,yk) кластердегі ұпайлардың орташа мәні болып табылады.

$$\bar{x}_k = \frac{1}{n_k} \sum_{i \in \text{кластер } k} x_i;$$

$$\bar{y}_k = \frac{1}{n_k} \sum_{i \in \text{кластер } k} y_i.$$

Кластердегі квадраттардың қосындысы келесі формуламен беріледі:

$$SS_k = \sum_{i \in \text{кластер } k} \left(x_i - \bar{x}_k\right)^2 + \left(y_i - \bar{y}_k\right)^2.$$

K-means әдісі барлық төрт кластердегі квадраттардың кластер ішіндегі қосындысын азайтатын жазбалардың кластерлерге тағайындалуын табады.

SS1+ SS2+ SS3+ SS4

$$\sum_{k=1}^{4} SS_i.$$

K-кластерлеуді оның кластерге бейімділігіне қатысты акциялар бағасының мінез-құлқын жақсы түсіну үшін пайдалануға болады. Акция кірістері дерлік стандартталған түрде хабарланатынын ескеріңіз, сондықтан деректерді қалыпқа келтірудің қажеті жоқ. R тілінде K-мағынасын кластерлеу kmeans функциясын пайдаланып орындалуы мүмкін. Мысалы, келесі код үзіндісі екі айнымалыға негізделген төрт кластерді табады, ExxonMobil (XOM) және Chevron (CVX) қор қайтарымы:

```
df <- sp500_px[row.names(sp500_px)>='2011-01-01', c('XOM', 'CVX')]
 km <- kmeans(df, centers=4)
```

Әрбір жазбаның кластеріне тағайындау кластер компонентінде қайтарылады:

```
> df$cluster <- factor(km$cluster)
 > head(df)
XOM        CVX cluster
2011-01-03 0.73680496  0.2406809       2
2011-01-04 0.16866845 -0.5845157       1
2011-01-05 0.02663055  0.4469854       2
2011-01-06 0.24855834 -0.9197513       1
2011-01-07 0.33732892  0.1805111       2
2011-01-10 0.00000000 -0.4641675       1
```

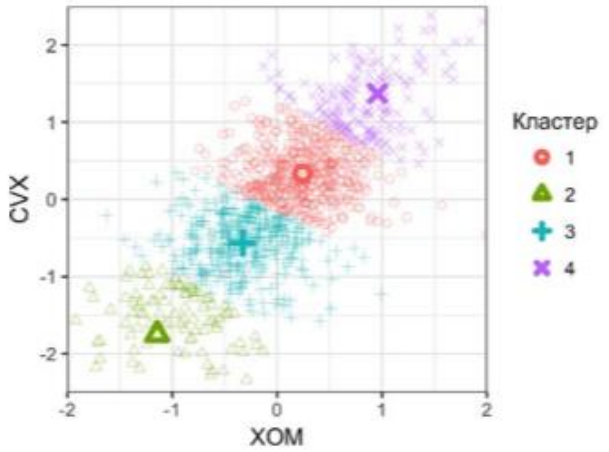
Алғашқы алты жазба 1 кластерге немесе 2 кластерге тағайындалады. Осы кластерлердің орталықтары (орташалары) да қайтарылады:

```
> centers <- data.frame(cluster=factor(1:4), km$centers) >
centers
cluster
XOM        CVX 1
1 -0.3284864 -0.5669135 2
2  0.2410159  0.3342130 3
3 -1.1439800 -1.7502975 4
4  0.9568628  1.3708892
```

1 және 3 кластерлер «құлайды» нарықтарды, ал 2 және 4 кластерлер «өсу» нарықтарын білдіреді. Бұл мысалда екі айнымалысы бар кластерлер мен олардың орталықтарының визуализациясы өте қарапайым:

ggplot(data=df, aes(x=XOM, y=CVX, color=cluster, shape=cluster)) + geom_point(alpha=.3) +
geom_point(data=centers, aes(x=XOM, y=CVX), size=3, stroke=2)

Алынған график 7.4 -суретте көрсетілген. кластер тапсырмаларын және кластер орталықтарын көрсетеді.



7.4-сурет. ExxonMobil және Chevron акцияларының бағасы туралы деректерге қолданылатын K кластерлері
(тығыз аймақтағы екі кластердің орталықтарын ажырату қиын)

K алгоритмі білдіреді
Жалпы, K алгоритмін p айнымалылары бар деректер жиынына қолдануға болатынын білдіреді
X1, ..., Xp. K -ортасының нақты шешімін табу есептеу тұрғысынан өте қиын болғанымен, эвристикалық алгоритмдерді қолдану арқылы жергілікті оңтайлы шешімді тиімді есептеуге болады.

Алгоритм пайдаланушы анықтаған K санынан және кластер орталықтарының бастапқы жинағынан басталады, содан кейін келесі қадамдар арқылы қайталанады:
1. Әрбір жазбаны өлшенген квадраттық қашықтыққа сәйкес ең жақын кластер орталығына тағайындаңыз.
2. Жазбаларды кластерлерге тағайындау негізінде кластерлердің орталықтарын жаңа әдіспен қайта есептеңіз.

Алгоритм кластерлерге жазбаларды тағайындау өзгермегенде жинақталады. Бірінші итерация үшін кластер орталықтарының бастапқы жинағын көрсету керек. Бұл әдетте әрбір жазбаны K кластерлерінің біріне кездейсоқ тағайындау және содан кейін осы кластерлердің орташа мәндерін табу арқылы жасалады. Бұл алгоритм ең жақсы шешімді табуға кепілдік бермейтіндіктен, алгоритмді инициализациялау үшін әртүрлі кездейсоқ үлгілерді пайдаланып алгоритмді бірнеше рет орындау ұсынылады. Бірнеше итерациялар жиыны пайдаланылғанда, K орташасының нәтижесі

квадраттардың ең төменгі кластерлік қосындысы бар итерация арқылы беріледі. R kmeans функциясының nstart параметрі кездейсоқ іске қосулар санын анықтауға мүмкіндік береді. Мысалы, келесі код үзіндісі 10 түрлі бастапқы кластер орталықтарын пайдаланып 5 кластерді табу үшін K ортасының алгоритмін орындайды:

```
syms <- c( 'AAPL', 'MSFT', 'CSCO', 'INTC', 'CVX', 'XOM', 'SLB', 'COP', 'JPM', 'WFC', 'USB', 'AXP', 'WMT', 'TGT', 'HD', 'COST')
df <- sp500_px[row.names(sp500_px)>='2011-01-01', syms]
km <- kmeans(df, centers=5, nstart=10)
```

Функция 10 түрлі бастапқы нүктеден ең жақсы шешімді автоматты түрде қайтарады. Әрбір кездейсоқ іске қосу үшін алгоритмге берілген итерациялардың ең көп санын анықтау үшін iter.max параметрін пайдалануға болады.

Кластерлік интерпретация

Кластерлік талдаудың маңызды бөлігі кластерлерді түсіндірумен байланысты болуы мүмкін. Кмеаннан шығатын ең маңызды екі деректер элементі кластер өлшемдері мен кластер орталықтары болып табылады. Мысалы, алдыңғы ішкі бөлімде алынған кластерлердің өлшемдері келесі R пәрменімен беріледі:

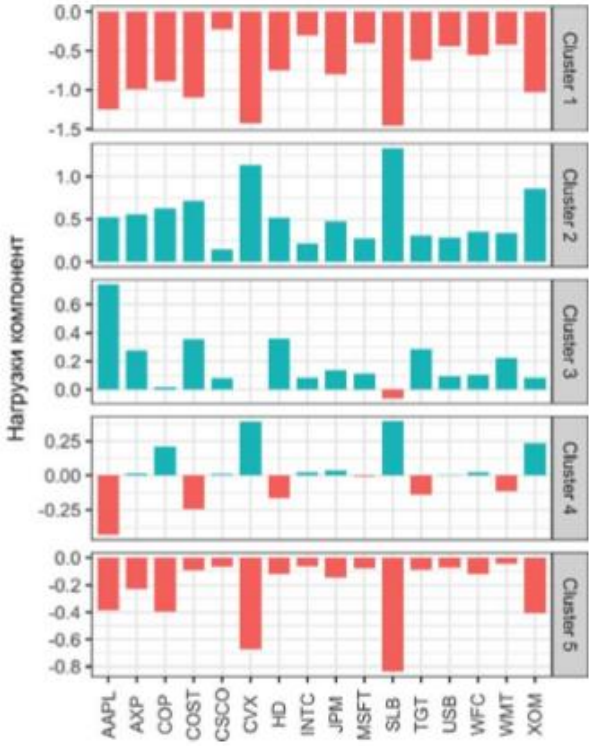
```
km$size
[1] 186 106 285 288 266
```

Кластер өлшемдері салыстырмалы түрде теңдестірілген. Теңгерімсіз кластерлер басқа деректерден өте ерекшеленетін алыс шеткі мәндерден немесе жазбалар топтарынан туындауы мүмкін, олардың екеуі де қосымша зерттеуді қажет етуі мүмкін. Сіз ggplot-пен бірге жинау функциясын пайдаланып кластер орталықтарын сыза аласыз:

```
centers <- as.data.frame(t(centers))
names(centers) <- paste("Cluster", 1:5)
 centers$Symbol <- row.names(centers)
centers <- gather(centers, "Cluster", "Mean", -Symbol)
centers$Color = centers$Mean > 0
ggplot(centers, aes(x=Symbol, y=Mean, fill=Color)) +
 geom_bar(stat='identity', position = "identity", width=.75) +
facet_grid(Cluster ~ ., scales='free_y')
```

Алынған график 7.5 суретте көрсетілген. және әрбір кластердің сипатын көрсетеді. Мысалы, 1 және 2 кластерлер нарық құлдырап, көтерілетін күндерге сәйкес келеді. 3 және 5 кластерлер тұтынушылық қорлар нарығының өсу күндерімен және энергия қорлары нарығының құлдырау

күндерімен сипатталады. Соңында, 4-кластер энергия қоры өскен және тұтыну нарығының қорлары төмендеген күндерді қамтиды.



7.5.-сурет. Әрбір кластердегі айнымалылардың орташа мәндері («центроидтар»)

Кластер санын таңдау

K мағыналарының алгоритмі K кластерлерінің санын анықтауды талап етеді. Кейде кластерлердің саны қолданбаға тән. Мысалы, сату ұйымы тұтынушыларды «түрлерге» топтастыруды және оларды коммерциялық ұсыныстармен шақыруды қалауы мүмкін. Мұндай жағдайда ұйымдық ойлар жоспарланған тұтынушылар сегменттерінің санын белгілейді — мысалы, екі сегмент тұтынушылардың пайдалы бөлінуіне әкелмеуі мүмкін, ал сегізі өңдеуге тым көп болуы мүмкін.

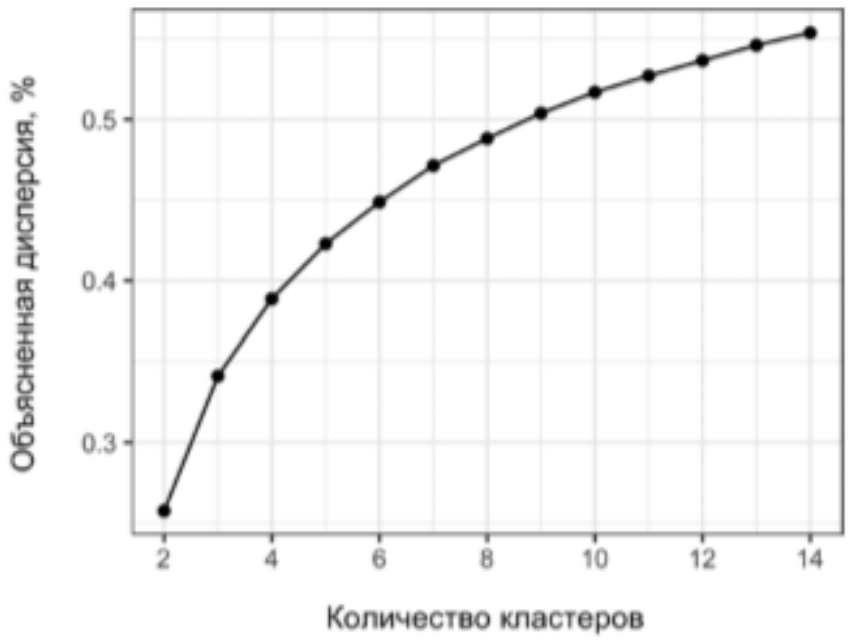
Тәжірибелік немесе ұйымдастырушылық себептерге байланысты кластерлердің саны болмаған жағдайда статистикалық тәсілді қолдануға болады. Кластерлердің «ең жақсы» санын табудың бірыңғай стандартты әдісі жоқ.

Шынтақ әдісі деп аталатын жалпы тәсіл кластерлер жинағы деректердегі дисперсияның «көпшілігін» түсіндіретін нүктені анықтау болып табылады. Осы жиыннан тыс жаңа кластерлерді қосу түсіндірілетін дисперсияға салыстырмалы түрде шағын қосымша үлес қосады. Шынтақ – тік көтерілуден

кейін жинақталған түсіндірілетін дисперсия деңгейі төмендейтін нүкте, сондықтан бұл әдістің атауы.

7.6-суретте 2-ден 15-ке дейінгі кластерлер саны үшін қайтарылмайтын деректер үшін түсіндірілген дисперсияның жиынтық пайызы көрсетілген. Бұл мысалдағы шынтақ қай жерде? Айқын үміткер жоқ, өйткені дисперсияның ұлғаюы бірте-бірте төмендейді. Бұл жақсы анықталған кластерлері жоқ деректер үшін әдеттегі жағдай. Мүмкін, бұл шынтақ әдісінің кемшілігі, бірақ ол деректердің сипатын көрсетеді. R тілінде kmeans функциясы шынтақ әдісін қолдану үшін бір пәрменді қамтамасыз етпейді, бірақ оны төменде көрсетілгендей кмеан шығысына негізделген оңай қолдануға болады.

```
pct_var <- data.frame(pct_var = 0,
num_clusters=2:14)
totalss <- kmeans(df, centers=14, nstart=50, iter.max = 100)$totss
for(i in 2:14){
pct_var[i-1, 'pct_var'] <- kmeans(df, centers=i, nstart=50, iter.max = 100)
$betweenss/totalss
```



7.6.-сурет Акция деректеріне қолданылатын шынтақ әдісі

Қанша кластер қалуы керек деп есептегенде, ең маңызды тексеру мынада: кластерлердің жаңа деректерде қайталану ықтималдығы қандай? Кластерлер интерпретацияланады және деректердің жалпы сипаттамаларына қатысты ма, әлде олар жай ғана белгілі бір дананы көрсете ме? Сіз мұны айқас валидация арқылы ішінара квалификациялауға болады.

Тұтастай алғанда, қанша кластер жасау керектігін сенімді түрде нұсқайтын бірде-бір ереже жоқ.

K кластерлеуге арналған негізгі идеялар:

- *Қажетті кластерлердің саны K пайдаланушы таңдайды.*
- *Алгоритм кластер тағайындаулары өзгермейінше, ең жақын кластер орталығына жазбаларды итеративті түрде тағайындау арқылы кластерлерді нақтылайды.*
- *K таңдауда әдетте практикалық ойлар басым болады; кластерлердің статистикалық анықталған оңтайлы саны жоқ.*

Иерархиялық кластерлеу

Иерархиялық кластерлеу әр түрлі кластерлерді құра алатын K-құралдарға балама әдіс болып табылады. Иерархиялық кластерлеу K мағынасына қарағанда икемді және сандық емес айнымалыларды қабылдау оңай. Ол қашықтағы немесе шектен тыс топтарды немесе жазбаларды анықтауда аса сезімтал. Иерархиялық кластерлеу интуитивті графикалық дисплейге де мүмкіндік береді, бұл кластерлерді оңай түсіндіруге әкеледі.

Негізгі терминдері:

Дендрограмма

Жазбалардың көрнекі көрінісі және олар тиесілі кластерлердің иерархиясы.

Қашықтық

Бір жазбаның екіншісіне қаншалықты жақын екенін көрсететін көрсеткіш.

Ұқсассыздық

Бір кластердің екіншісіне жақындық дәрежесінің метрикалық көрсеткіші.

Синонимдер: ұқсас емес, ұқсас емес.

Иерархиялық кластерлеудің икемділігі қымбатқа түседі — иерархиялық кластерлеу миллиондаған жазбалары бар үлкен деректер жиынына жақсы масштабталмайды. Тіпті он мыңдаған жазбалары бар қарапайым өлшемді деректер үшін де иерархиялық кластерлеу есептеуді қарқынды болуы мүмкін. Шынында да, иерархиялық кластерлеу қосымшаларының көпшілігі салыстырмалы түрде шағын деректер жиынына бағытталған.

Қарапайым мысал

Иерархиялық кластерлеу n жазба және p айнымалысы бар деректер жиынында жұмыс істейді және екі негізгі құрылымдық блокқа негізделген:

метрикалық қашықтық
dij, ол арасындағы қашықтықты өлшейді
екі жазба і және j;

метрикалық айырмашылық
Dab, ол қашықтықтарға негізделген екі А және В кластері арасындағы айырмашылықты өлшейді , dij әрбір кластердің мүшелері арасындағы. Сандық деректерді қамтитын қолданбалар үшін ең маңызды шешім айырмашылық метрикасын таңдау болып табылады. Иерархиялық кластерлеу әрбір жазбаны өз кластері ретінде орнату арқылы басталады (яғни, біртұтас кластерлер жазбалар саны бойынша жасалады) және ең аз ұқсамайтын кластерлерді біріктіру үшін қайталанатын түрде орындалады.

R тілінде hclust функциясы иерархиялық кластерлеуді орындау үшін пайдаланылуы мүмкін. Hclust пен kmeans арасындағы үлкен айырмашылық мынада, бұл функция деректердің өзінде емес, ijd жұптық қашықтықта жұмыс істейді. Сіз оларды dist функциясы арқылы есептей аласыз. Мысалы, келесі код үзіндісі бірқатар компаниялардың акцияларының кірістеріне иерархиялық кластерлеуді қолданады:

```
syms1 <- c('GOOGL', 'AMZN', 'AAPL', 'MSFT', 'CSCO', 'INTC', 'CVX',
'XOM', 'SLB', 'COP', 'JPM', 'WFC', 'USB', 'AXP',
 'WMT', 'TGT', 'HD', 'COST')
# транспонировать: чтобы кластеризовать компании, нужно
# разместить акции в строках
df <- t(sp500_px[row.names(sp500_px)>='2011-01-01', syms1])
d <- dist(df)
hcl <- hclust(d)
```

Кластерлеу алгоритмі деректер кадрының жазбаларын (жолдарын) кластерлерге таратады. Біз компанияларды кластерлеуді қалайтындықтан, деректер кадрын ауыстырып, қорларды жолдар мен күндерді бағандарға таратуымыз керек.

Дендограмма

Иерархиялық кластерлеу дендограмма деп аталатын ағаш түріндегі табиғи графикалық бейнелеуге мүмкіндік береді. Бұл атау гректің dendro (ағаш) және грамма (сызу) сөздерінен шыққан. R тілінде plot пәрменін пайдаланып дендрограмманы оңай жасауға болады:

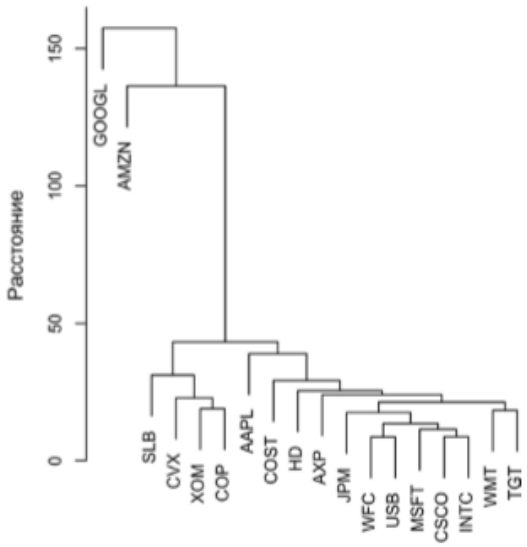
```
сюжет(hcl)
```

Нәтиже күріште көрсетілген. 7.7. Ағаштың жапырақтары жазбаларға сәйкес келеді. Ағаш бұтақтарының ұзындығы сәйкес кластерлер арасындағы айырмашылық дәрежесін көрсетеді. Google және Amazon акцияларының

кірістері басқа акциялардың кірістерінен айтарлықтай ерекшеленеді. Басқа қорлар табиғи топтарға бөлінеді: энергия қорлары, қаржылық қорлар және тұтыну секторы қорлары, барлығы өздерінің ішкі ағаштарында орналастырылған. K білдіреді әдісінен айырмашылығы, кластерлердің санын алдын ала анықтаудың қажеті жоқ. Кластерлердің белгілі бір санын шығару үшін cutree функциясын пайдалануға болады:

```
cutree(hcl, k=4)
GOOGL AMZN AAPL MSFT CSCO INTC CVX XOM SLB COP JPM WFC
1    2    3    3    3    3    4    4    4    4    3    3
 USB  AXP  WMT  TGT   HD COST
3    3    3    3    3    3
```

Шығарылатын кластерлердің саны 4 деп есептеледі және Google және Amazon бөлісулерінің әрқайсысы өз кластеріне тиесілі екенін көре аласыз. Мұнай акциялары (XOM, CVS, SLB, COP) басқа кластерге жатады. Қалған акциялар төртінші кластерде.



7.7-сурет. Қор дендограммасы
(Дендограмма акция)

Агломеративті алгоритм

Негізгі иерархиялық кластерлеу алгоритмі ұқсас кластерлерді итеративті түрде біріктіретін агломеративті алгоритм болып табылады. Агломеративті алгоритм әрбір жазбаны өзінің синглтондық кластерін жасаудан бастайды, содан кейін үлкенірек және үлкенірек кластерлерді құрады. Бірінші қадам - жазбалардың барлық жұптары арасындағы қашықтықты есептеу.

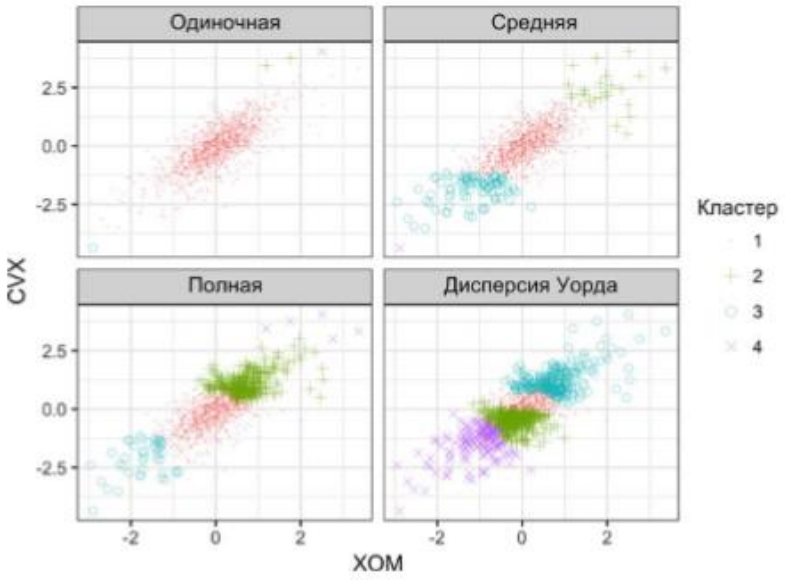
Айырмашылық өлшемдері

Айырмашылықтың төрт жалпы қабылданған өлшемі бар: толық сілтеме, жалғыз сілтеме, орташа сілтеме және минималды дисперсия. Жоғарыда аталғандардың барлығына (басқа шараларға қоса) hclust бағдарламалық пакетін қоса алғанда, иерархиялық кластерлік бағдарламалық жасақтама жүйелерінің көпшілігі қолдау көрсетеді. Бұрын анықталған толық байланыс әдісі ұқсас мүшелері бар кластерлерді жасайды. Жалғыз сілтеме әдісі екі кластердегі жазбалар арасындағы қашықтықты азайтады:

$$D(A, B) = \min d(a_i, b_j) \text{ для всех пар } i, j.$$

Бұл «ашкөз» әдіс және ол мүлдем басқа элементтерді қамтитын кластерлерді шығарады. Орташа қосылу әдісі қарастырылып отырған жұптардың барлық арақашықтықтарының орташа мәні болып табылады және ол жалғыз және толық қосылу әдістерінің арасындағы ымыраны білдіреді. Соңында, Ward әдісі деп те аталатын минималды дисперсия әдісі K-орташа әдісіне ұқсас, өйткені ол квадраттардың кластерлік қосындысын азайтады.

7.8-суретте ExxonMobil және Chevron акциялары бойынша кірістердің төрт өлшемін қолданатын иерархиялық кластерлеу көрсетілген. Әрбір өлшем үшін төрт кластер қалдырылады. Нәтижелер таң қалдырады: бір сілтеме өлшемі барлық дерлік нүктелерді бір кластерге орналастырады. Минималды дисперсия әдісін (Уорд әдісі) қоспағанда, барлық өлшемдер жалпы алғанда, бірнеше шеткі нүктелері бар кем дегенде бір кластермен аяқталады. Минималды дисперсия әдісі К ортасының кластеріне ең жақын; 7.4.-суретпен салыстыру.



7.8.-сурет. Қор деректеріне қолданылатын айырмашылық өлшемдерін салыстыру

Иерархиялық кластерлеудің негізгі идеялары:

- *Әрбір жазбаны өз кластеріне қоюдан бастау керек.*
- *Барлық жазбалар бір кластерге (агломеративті алгоритм) тиесілі болғанша кластерлер көрші кластерлермен біртіндеп біріктіріледі.*
- *Агломерация тізбегі сақталады және графикте белгіленеді, ал пайдаланушы (кластерлердің санын алдын ала көрсетпей) әртүрлі кезеңдердегі кластерлердің саны мен құрылымын көрнекі түрде көре алады.*
- *Кластераралық қашықтық әртүрлі тәсілдермен есептеледі, олардың барлығы жазбалар арасындағы барлық қашықтықтардың жиынтығына сүйенеді.*