Модельге негізделген кластерлеу

Иерархиялық кластерлеу және K-орталары сияқты кластерлеу әдістері эвристикаға негізделген және ең алдымен деректерден тікелей алынған өлшемдерге сәйкес мүшелері бір-біріне жақын кластерлерді табуға негізделеді (ықтималдық модель тартылмайды). Соңғы 20 жылда модельге негізделген кластерлеу әдістерін әзірлеуге көп күш жұмсалды. Эдриан Рафтери және Вашингтон университетінің басқа зерттеушілері теория мен бағдарламалық жасақтаманы қоса алғанда, модельге негізделген кластерлеу әдістеріне елеулі үлес қосты. Бұл әдістер статистикалық теорияға сүйенеді және кластерлердің сипаты мен санын анықтаудың неғұрлым қатаң жолдарын ұсынады. Олар, мысалы, бір-біріне ұқсас, бірақ міндетті түрде бір-біріне жақын емес жазбалардың бір тобы болуы мүмкін жағдайларда (мысалы, кірістердің дисперсиясы жоғары технологиялық қорлар) және ұқсас және жақын орналасқандарды (мысалы, дисперстілігі төмен коммуналдық қорлар) жазады.

Көп өлшемді қалыпты таралу

Ең көп қолданылатын модельге негізделген кластерлеу әдістері көп айнымалы қалыпты үлестірімге сүйенеді. Бұл p айнымалылар X , X2 , ..., X p жиынына қалыпты үлестірудің жалпылауы. Бөлу μ = μ1, μ2 , ..., μp және коварианттық матрицасы арқылы анықталады. Σ. Коварианттық матрицасы айнымалылардың бір-бірімен корреляциясының өлшемі болып табылады (ковариация туралы қосымша ақпаратты 5-тараудың «Коварианттық матрицасы» бөлімінен қараңыз). Σ коварианттық матрицасы p дисперсияларынан σ квадрат 1, σ квадрат 2, ..., σ квадрат p және i ≠ j айнымалылардың барлық жұптары үшін ковариациялардан , σ ij тұрады. Жолдар бойынша таратылған және бағандар бойынша қайталанатын айнымалы мәндермен матрица келесідей болады:

$$\mathbf{\Sigma}=\begin{bmatrix}\sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p}\\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,p}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_p^2\end{bmatrix}.$$

Ковариация матрицасы симметриялы болғандықтан, және σ i, j =σ j, i тек бар p× (p -1) -p ковариацияның мүшелері. Барлығы коварианттық матрицада p× (p -1) параметрлері бар. Бөлу келесідей белгіленеді:

$$\left(X_1, X_2, ..., X_p\right)\tilde{N}_p\left(\mu, \mathbf{\Sigma}\right).$$

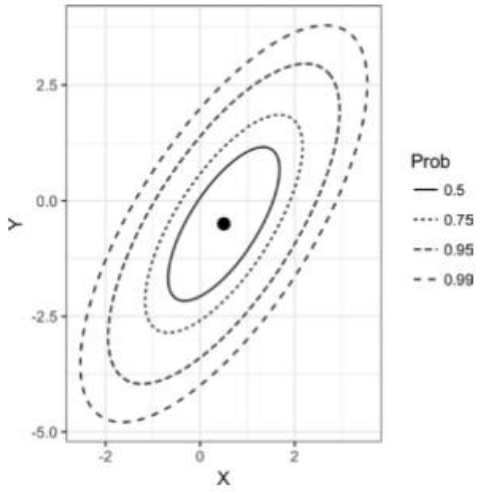
Бұл барлық айнымалылар қалыпты таралған және жалпы таралу айнымалылардың орташа векторы мен коварианттық матрица арқылы толығымен сипатталатынын айтудың аналитикалық тәсілі.

Суретте. 7.9-суретте екі X және Y айнымалысының көп айнымалы қалыпты таралуы үшін ықтималдық контурлары көрсетілген (0,5 ықтималдық контуры, мысалы, үлестірімнің 50% қамтиды).

Орташа мәндер μ x = 0,5 және μy =− 0,5 , ал коварианттық матрицасы:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}.$$

σ xy ковариациясы оң болғандықтан, X және Y оң корреляцияланады:



7.9. –сурет. Екі айнымалы қалыпты таралу үшін ықтималдық контурлары

Қалыпты таралу қоспалары

Модельге негізделген кластерлеу әрбір жазба K көп айнымалы қалыпты таралулардың бірі ретінде таратылады деген негізгі идеяға негізделген, мұнда K - кластерлердің саны. Әрбір үлестірімнің әртүрлі орташа μ және коварианттық матрицасы Σ болады. Мысалы, егер X және Y екі айнымалы болса, онда әрбір жол (Xi,Yi) K үлестірімдерінің (N1(μ1), Σ1), (N2(μ2), Σ2),.. (N1(μK),ΣK) бірінен таңдалғандай модельденеді. R-де Модельге негізделген кластерлеуге арналған mclust деп аталатын өте бай бағдарламалық пакет бар, оны бастапқыда Крис Фрейли мен Адриан Рафтери әзірлеген. Осы пакеттің көмегімен біз K-Means және иерархиялық кластерлеу

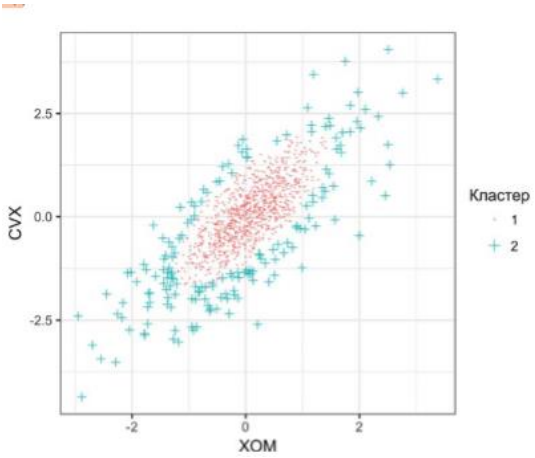
әдістерін қолдана отырып, бұрын талдаған қор қайтарымы деректеріне Үлгіге негізделген кластерлеуді қолдана аламыз:

```
> library(mclust)
> df <- sp500_px[row.names(sp500_px)>='2011-01-01', c('XOM', 'CVX')]
 > mcl <- Mclust(df)
 > summary(mcl)
Mclust VEE (ellipsoidal, equal shape and orientation) model with 2 components:
log.likelihood    n df       BIC       ICL
 -2255.134 1131  9 -4573.546 -5076.856
Clustering table:
    1   2
    963 168
```

Осы код бөлігін іске қоссаңыз, есептеу басқа процедураларға қарағанда айтарлықтай ұзағырақ болатынын байқайсыз. Болжау функциясы арқылы кластер тапсырмаларын шығарғаннан кейін біз кластерлерді көрнекі түрде көрсете аламыз:

```
cluster <- factor(predict(mcl)$classification)
ggplot(data=df, aes(x=XOM, y=CVX, color=cluster, shape=cluster)) + geom_point(alpha=.8)
```

Алынған график 7.10. суретте көрсетілген. Екі кластер бар: бір кластер деректердің ортасында және екінші кластер деректердің сыртқы жиегінде. Бұл ықшам пішіні бар кластерлерді табатын K құралдары (7.4-суретті қараңыз) және иерархиялық кластерлеу (7.8-суретті қараңыз) көмегімен алынған кластерлерден өте ерекшеленеді.



7.10.-сурет. mclust бумасын пайдаланып акция қайтару деректері үшін алынған екі кластер

summary функциясының көмегімен қалыпты үлестірімнің параметрлерін шығара аламыз:

> summary(mcl, parameters=TRUE)$mean
[,1] [,2]
XOM 0.05783847 -0.04374944
CVX 0.07363239 -0.21175715
> summary(mcl, parameters=TRUE)$variance
, , 1
XOM CVX
XOM 0.3002049 0.3060989 CVX 0.3060989 0.5496727 , , 2
XOM CVX XOM 1.046318 1.066860
CVX 1.066860 1.915799

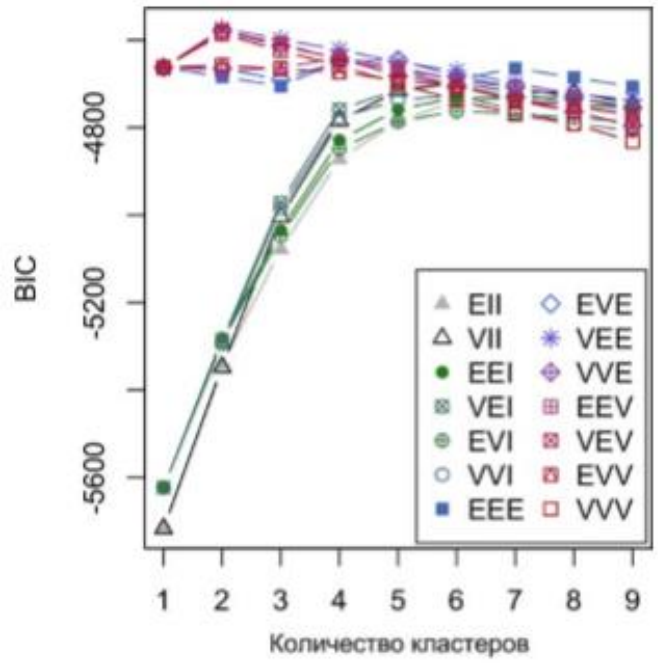
Бөлулердің ұқсас құралдары мен корреляциялары бар, бірақ екінші таралуда әлдеқайда үлкен дисперсиялар мен ковариациялар бар. mclust кластерлері таңқаларлық болып көрінуі мүмкін, бірақ іс жүзінде олар әдістің статистикалық сипатын көрсетеді. Модельге негізделген кластерлеудің мақсаты оңтайлы бекітілген көп айнымалы қалыпты үлестірулер жинағын табу болып табылады. Қор деректері бетінде қалыпты пішінге ие болып көрінеді (7.9-суреттегі контурларды қараңыз). Шын мәнінде, акциялардың кірістері қалыпты үлестірімге қарағанда ұзағырақ таралуға ие. Бұны өңдеу үшін mclust таратуды деректердің негізгі бөлігіне сәйкестендіреді, бірақ одан кейін дисперсиясы көбірек екінші таратуға сәйкес келеді.

Кластер санын таңдау

K құралдары мен иерархиялық кластерлеуден айырмашылығы, mclust кластерлердің санын автоматты түрде таңдайды (бұл жағдайда екі). Ол мұны Байес ақпараттық критерийлері (BIC) ең жоғары мәнге ие кластерлердің санын таңдау арқылы жасайды. BIC (AIC-ге ұқсас) – мүмкін болатын модельдер жиынтығының ішінен ең жақсы үлгіні табуға арналған жалпы құрал. Мысалы, AIC (немесе BIC) қадамдық регрессияда үлгі таңдау үшін кеңінен қолданылады (4-тараудағы «Модельді таңдау және қадамдық регрессия» бөлімін қараңыз). BIC модельдегі параметрлер саны үшін айыппұлмен ең қолайлы үлгіні таңдау арқылы жұмыс істейді. Үлгіге негізделген кластерлеу жағдайында көбірек кластерлерді қосу үлгіге қосымша параметрлерді енгізу арқылы сәйкестікті әрқашан жақсартады. Hclust ішіндегі функцияны пайдаланып әрбір кластер өлшемі үшін BIC мәндерін салуға болады:

сюжет(mcl, не='BIC', сұрау=ЖАЛҒАН)

Кластерлердің саны — немесе әртүрлі көп өлшемді қалыпты үлгілердің (компоненттер) саны — x осінде көрсетілген (7.11-сурет). Бұл сызба K құралдары үшін кластерлердің санын анықтау үшін пайдаланылатын шынтақ сызбасына ұқсайды, тек сызба түсіндірілген дисперсия пайызынан гөрі BIC мәнін көрсетеді (7.6-суретті қараңыз). Бір үлкен айырмашылық, бір жолдың орнына mclust 14 түрлі жолды көрсетеді! Себебі, mclust әрбір кластер өлшемі үшін 14 түрлі үлгіні сәйкестендіруді орындайды және соңында ең қолайлы үлгіні таңдайды.



7.11.-сурет. Әртүрлі кластерлер саны (компонент) үшін қор қайтарымы деректері үшін BIC бағалаулары

Неліктен mclust көп айнымалы қалыпты үлестірулердің ең жақсы жинағын анықтау үшін көптеген үлгілерге сәйкес келеді? Өйткені модельге сәйкес келетін коварианттық матрицаны Σ параметрлеудің әртүрлі тәсілдері бар. Көп жағдайда үлгілердің егжей-тегжейлері туралы алаңдамаудың қажеті жоқ және сіз mclust бумасы таңдаған үлгіні қауіпсіз пайдалана аласыз. Бұл мысалда, BIC сәйкес, үш түрлі модель (VEE, VEV және VVE деп аталады) екі құрамдастың көмегімен оңтайлы сәйкестікті береді.

Үлгіге негізделген кластерлеу әдістері, алайда, бірнеше шектеулер бар. Бұл әдістер деректерге арналған үлгінің сипаты туралы негізгі болжамды талап етеді және кластер нәтижелері осы болжамға өте тәуелді. Оның есептеу қажеттіліктері тіпті иерархиялық кластерлеуден де жоғары, бұл үлкен

деректерге масштабтауды қиындатады. Ақырында, оның алгоритмі басқа әдістерге қарағанда күрделірек және қол жетімді емес.

Үлгіге негізделген кластерлеуге арналған негізгі идеялар:

- *Кластерлер әртүрлі ықтималдық үлестірімдері бар әртүрлі деректерді генерациялау процестерінен келеді деп болжанады.*
- *Әрі қарай, әртүрлі (әдетте қалыпты) таралу сандарын ескере отырып, әртүрлі модельдік фитингтер орындалады.*
- *Бұл әдіс тым көп параметрлерді пайдаланбай (яғни, артық орнату) деректерге жақсы сәйкес келетін үлгіні (және онымен байланысты кластерлер санын) таңдайды.*

Масштабтау және категориялық айнымалылар

Бақыланбайтын оқыту әдістері әдетте деректердің сәйкес масштабта болуын талап етеді. Бұл масштабтау маңызды емес көптеген регрессия және жіктеу әдістерінен айырмашылығы (ерекшелік - K ең жақын көршілер әдісі).

Негізгі терминдері:

- *Масштабтау Деректерді тегістеу немесе кеңейту, әдетте бірнеше айнымалыларды бірдей өлшем шкаласына келтіру үшін.*
- *Нормалау Масштабтау әдістерінің бірі орташа мәнді алып тастау және стандартты ауытқуға бөлу болып табылады. Синонимі: стандарттау.*
- *Гоуэр қашықтығы Барлық айнымалы мәндерді 0-1 аралығына келтіру үшін аралас сандық және категориялық деректерге қолданылатын масштабтау алгоритмі.*

Мысалы, жеке несие деректері туралы айтқанда, айнымалылар бірліктер мен мәндердің алуан түрлілігіне ие. Кейбір айнымалылар салыстырмалы түрде шағын мәндерге ие (мысалы, пайдаланылған жылдар саны), ал басқалары өте үлкен мәндерге ие (мысалы, несиенің доллар сомасы). Егер деректер масштабталмаған болса, онда үлкен мәндері бар айнымалылар PCA, K құралдары және басқа кластерлеу әдістерінен басым болады, ал шағын мәндері бар айнымалылар олармен еленбейді. Кейбір кластерлеу процедуралары үшін категориялық деректер белгілі бір мәселе болуы мүмкін.

Ең жақын көршілес K сияқты, ретсіз факторлық айнымалылар әдетте бір белсенді күйді кодтау арқылы екілік (0/1) айнымалылар жиынына түрлендіріледі (6-тараудың Бір белсенді күйді кодтаушы бөлімін қараңыз). Екілік айнымалылардың басқа деректер шкалаларына қарағанда өлшеу шкаласы басқаша, сонымен қатар екілік айнымалылардың тек екі мәні бар екендігі PCA және K құралдары сияқты әдістермен проблемалар тудыруы мүмкін.

Доминант айнымалылар

Айнымалылар бірдей өлшем шкаласында және салыстырмалы маңыздылығын дәл көрсететін жағдайларда да (мысалы, акциялар бағасының динамикасы), кейде айнымалы мәндерді қайта масштабтау пайдалы болуы мүмкін. Бөлімде талдауға Alphabet (GOOGL) және Amazon (AMZN) қостық делік. Осы тараудың басында «Негізгі компоненттерді түсіндіру».

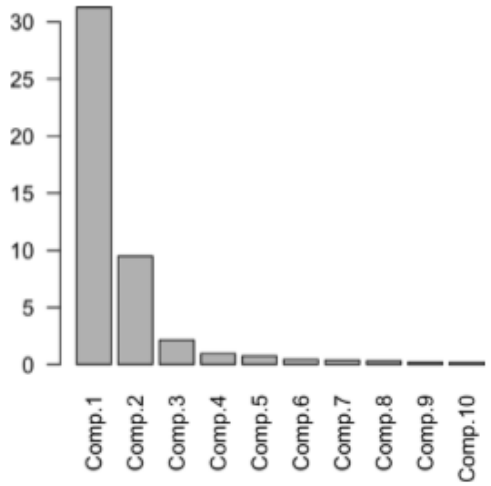
```
syms <- c('AMZN', 'GOOGL' 'AAPL', 'MSFT', 'CSCO', 'INTC', 'CVX', 'XOM',
'SLB', 'COP', 'JPM', 'WFC', 'USB', 'AXP', 'WMT', 'TGT', 'HD', 'COST')
top_sp1 <- sp500_px[row.names(sp500_px)>='2005-01-01', syms]
 sp_pca1 <- princomp(top_sp1)
screeplot(sp_pca1)
```

Скрипт сызбасы жоғарғы негізгі құрамдастардың ауытқуларын көрсетеді. Бұл жағдайда сызба сызбасы 7.12-суретте бірінші және екінші құрамдастардың дисперсиялары басқаларына қарағанда әлдеқайда үлкен екенін көрсетеді. Бұл көбінесе бір немесе екі айнымалының жүктемелерде басым екенін көрсетеді. Бұл мысалда солай:

```
round(sp_pca1$loadings[,1:2], 3)
Comp.1 Comp.2 GOOGL  0.781  0.609
AMZN   0.593 -0.792
 AAPL   0.078  0.004
MSFT   0.029  0.002
 CSCO   0.017 -0.001
 INTC   0.020 -0.001
CVX    0.068 -0.021
XOM    0.053 -0.005
```

Алғашқы екі негізгі құрамдас толығымен дерлік GOOGL және AMZN басым. Себебі GOOGL және AMZN акцияларының бағалары құбылмалылықта

басым. Бұл жағдайды шешу үшін оларды сол күйінде қосуға, айнымалы мәндерді қайта масштабтауға (осы тараудың басындағы «Айнымалыларды масштабтауды» қараңыз) немесе басым айнымалыларды талдаудан шығарып, оларды бөлек өңдеуге болады. «Дұрыс» тәсіл жоқ және шешім қолданбаға байланысты.



7.12.-сурет. GOOGL және AMZN қоса алғанда, S&P 500 бойынша PCA акцияларының кірістері үшін скрин диаграммасы

Категориялық деректер және басқару қашықтығы

Категориялық деректер жағдайында оны сандық деректерге рейтинг арқылы (реттік фактор үшін) немесе екілік (жалған) айнымалылар жиыны ретінде кодтау арқылы түрлендіру керек. Егер деректер аралас үздіксіз және екілік айнымалылардан тұратын болса, әдетте ауқымдар ұқсас болатындай айнымалы мәндерді қайта масштабтау қажет болады (осы тараудың басындағы «Айнымалы мәндерді масштабтау» бөлімін қараңыз). Танымал әдістердің бірі - Hower қашықтықты пайдалану. Гоуэрдің қашықтығы деректер түріне байланысты әр айнымалыға әртүрлі қашықтық метрикасын қолдану идеясына негізделген:

- сандық айнымалылар мен реттік факторлар үшін қашықтық екі жазба арасындағы айырмашылықтың абсолютті мәні ретінде есептеледі (Манхэттен қашықтығы);
- категориялық айнымалылар үшін екі жазба арасындағы санаттар әртүрлі болса, қашықтық 1 және санаттар бірдей болса, 0 болады.

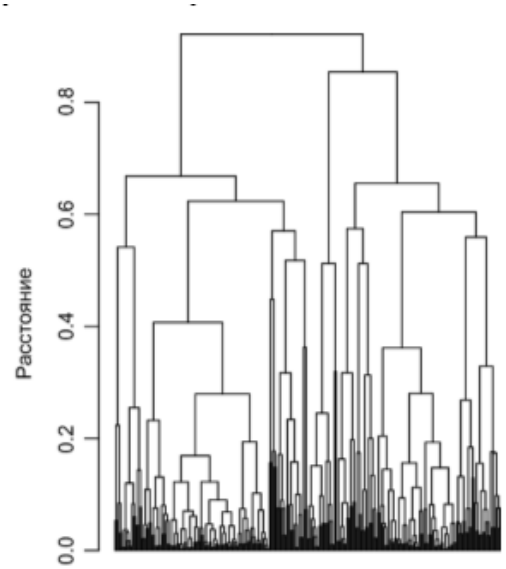
Говер қашықтығы келесідей есептеледі:

1.Қашықтықты есептеңіз , әрбір жазба үшін і және j айнымалылардың барлық жұптары үшін dij.
2.Әр жұпты масштабтаңыз , dij минимум 0 болуы үшін және максимум 1 болды.
3.Қашықтық матрицасын жасау үшін қарапайым немесе өлшенген орташа мәнді пайдаланып, айнымалылар арасындағы масштабталған қашықтықтарды жұптаңыз.

Алынған дендрограмма 7.13. суретте көрсетілген. Жеке жазбаларды x осінде ажыратуға болмайды, бірақ біз ішкі ағаштардың біріндегі жазбаларды (сол жақта, 0,5 «кесінді» пайдалану арқылы) келесі код үзіндісі арқылы тексере аламыз:

```
> df[labels(dnd_cut$lower[[1]]),]
# A tablle: 9 × 4
dti payment_inc_ratio  home   purpose   <dbl>
 <dbl>  <fctr>  <fctr>
1 24.57         0.83550    RENT   other
2 34.95         5.02763    RENT   other
3  1.51         2.97784    RENT   other
4  8.73        14.42070    RENT   other
5 12.05         9.96750    RENT   other
6 10.15        11.43180    RENT   other
7 19.61        14.04420    RENT   other
8 20.92         6.90123    RENT   other
9 22.49         9.36000    RENT   other
```

Бұл ішкі ағаш толығымен «басқа» деп белгіленген несие беру мақсатында жалға алушылардан тұрады. Барлық ішкі ағаштар үшін қатаң бөлу бар деп айту мүмкін болмаса да, график категориялық айнымалылардың кластерге бейімділігін көрсетеді.



7.13.-сурет. Аралас айнымалылар түрлерімен жұмыс істемейтін несиелер бойынша деректер үлгісіне қолданылған дендограмма hclust

Аралас мәліметтерді кластерлеу мәселелері

K-орталары және PCA әдістері үздіксіз айнымалылар үшін ең қолайлы. Кішірек деректер жинақтары үшін Гоуэр қашықтығымен иерархиялық кластерлеуді қолданған дұрыс. Негізінде, екілік немесе категориялық деректерге K құралдарын қолданбауға ешқандай себеп жоқ. Категориялық деректерді сандық мәндерге түрлендіру үшін әдетте «Бір белсенді кодтаушы» көрінісін (6-тараудың «Бір белсенді кодтаушы» бөлімін қараңыз) пайдаланасыз. Бірақ іс жүзінде екілік деректермен K-орталарын және PCA әдістерін пайдалану қиын болуы мүмкін.

Стандартты z-баллдары пайдаланылса, кластерлердің анықтамасында екілік айнымалылар басым болады. Себебі 0/1 пішіміндегі айнымалылар тек екі мәнді қабылдайды және K құралдары 0 немесе 1 бар барлық жазбаларды бір кластерге тағайындау арқылы квадраттардың шағын кластерлік сомасын ала алады. Мысалы, home және pub_rec_zero факторының айнымалы мәндерін қоса, үмітсіз несие деректеріне kmeans қолданайық:

```
df <- model.matrix(~ -1 + dti + payment_inc_ratio + home + pub_rec_zero,
                   data=defaults)
 df0 <- scale(df) km0 <-
kmeans(df0, centers=4, nstart=10)
centers0 <-scale(km0$centers, center=FALSE,
scale=1/attr(df0, 'scaled:scale'))
scale(centers0, center=-attr(df0, 'scaled:center'), scale=F)
dti payment_inc_ratio homeMORTGAGE homeOWN homeRENT pub_rec_zero
1 17.02            9.10        0.00       0     1.00         1.00
```

2 17.47	*8.43*	*1.00*	*0*	*0.00*	*1.00*
3 17.23	*9.28*	*0.00*	*1*	*0.00*	*0.92*
4 16.50	*8.09*	*0.52*	*0*	*0.48*	*0.00*

Үздік төрт кластер факторлық айнымалылардың әртүрлі деңгейлері үшін негізінен эрзат болып табылады. Бұл әрекетті болдырмау үшін екілік айнымалы мәндерді масштабтауға және басқа айнымалыларға қарағанда дисперсияны кішірек алуға болады. Немесе өте үлкен деректер жиындары үшін кластерлеуді нақты категориялық мәндерді қабылдайтын деректердің әртүрлі ішкі жиындарына қолдануға болады. Мысалы, ипотека төлейтін, үйді тікелей иеленетін немесе оны жалға алған жеке тұлғаларға берілген несиелерге кластерлеуді бөлек қолдануға болады.

***Деректерді масштабтауға арналған негізгі идеялар**:*

- *Әртүрлі өлшем шкалаларының айнымалылары олардың алгоритмдерге әсері негізінен өлшеу масштабымен анықталмайтындай етіп бір шкалаға келтірілуі керек.*
- *Жалпы масштабтау әдісі нормалау (стандарттау) болып табылады — орташа мәнді алып тастау және стандартты ауытқуға бөлу.*
- *Басқа әдіс, Гоуэр қашықтығы, барлық айнымалы мәндерді 0–1 диапазонына дейін масштабтайды (ол жиі аралас сандық және категориялық деректермен қолданылады).*