



Лекция 3:

Тематическое моделирование, метод главных компонент

Статистическое обучение без учителя

- Иногда называют задачей «самоорганизации»
- Все признаки равнозначны, нет отклика $X = (x_1, \dots, x_n)$, поэтому:
 - Обучение без учителя более «субъективное» (нет единых интуитивно понятных мер качества типа «точности») и менее «автоматизируемо»
 - Сложнее подбирать метапараметры и сравнивать полученные модели
- Важность этого направления в Data mining велико:
 - «Истинный data mining» - ищет неизвестные заранее зависимости без «подсказок» эксперта
 - Много важных прикладных задач по сегментации, выявление скрытых характеристик, зависимостей между атрибутами и т.д.
 - Для больших данных получить качественно размеченный набор тяжело или невозможно, часто возникают задачи *semisupervised learning*, когда размечена лишь часть набора

Основные задачи обучения без учителя

- Большинство задач сводится к поиску скрытых (латентных) признаков или скрытых структур (групп или зависимостей) в данных:

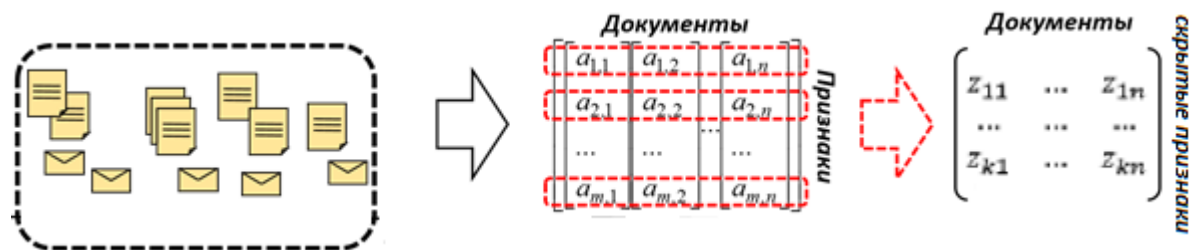
$$z_1 = F_1(x_1, \dots, x_n), \dots, z_k = F_k(x_1, \dots, x_n)$$

- Основные случаи: z или кластеры, или признаки, иногда как в SOM (сетях Кохонена) по сути и то, и другое.
- Основные прикладные задачи если z - признаки:
 - Предобработка данных (как правило перед применением регрессионных моделей или методов кластеризации) = сокращение пространства признаков $k \ll n$, удаление корреляций (или других зависимостей)
 - Тематическое моделирование текстовых данных (переход из пространства термов в пространство тематик)
 - Свертка «транзакционной истории» в вектор признаков, особенно часто используется в рекомендательных системах.

Тематическое моделирование текстов

■ Исходная задача:

- Каждый документ $(a_1, \dots, a_m)$ – вектор признаков в пространстве словаря, например созданный по схеме «мешок слов», a_i вес i -го слова в документе
- Корпус (множество документов) представим в виде матрицы



■ Тематическая модель:

- $z_1 = F_1(a_1, \dots, a_m), \dots, z_k = F_k(a_1, \dots, a_m)$, где $k \ll n$ - скрытые признаки документа или веса тематик

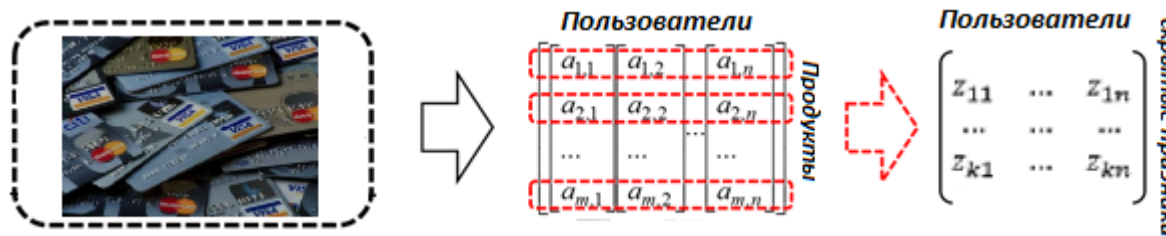
Рекомендательные системы

- Дано:
 - Множество пользователей (клиентов) J
 - Множество продуктов (услуг, ресурсов и т.д.) I
 - Множество транзакций D <время, пользователь, услуга>
- Задача:
 - Построить модель, которая по истории пользователя будет способна предсказать какие продукту (услуги или ресурсы) его заинтересуют.
- Решение на основе ассоциативных правил:
 - реализовано в узле Link Analysis
 - подставить события из истории пользователя во все правила и найти наиболее достоверное следующее событие.
- Альтернатива – на основе тематического моделирования

Рекомендательные системы

■ Исходная задача:

- Транзакционная история представляется в виде матрицы, где a_{ij} , например, равно числу транзакций пользователя j по использованию продукта i



■ Тематическая модель:

- $z_1 = F_1(a_1, \dots, a_m), \dots, z_k = F_k(a_1, \dots, a_m)$, где $k \ll n$ - скрытые предпочтения пользователя или веса тематик

Общая идея PCA

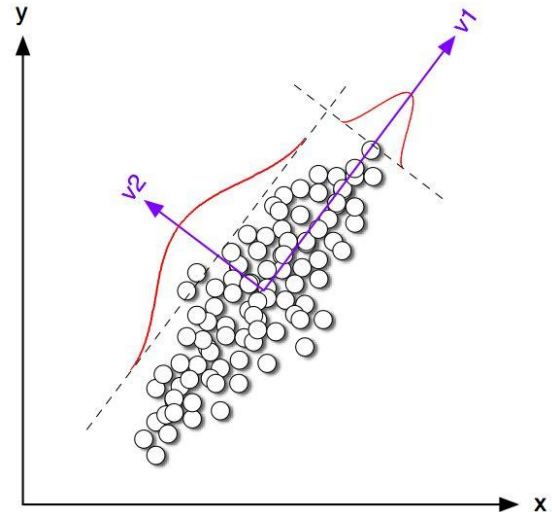
- Строится новый базис (линейное преобразование исходного пространства) такой, что:
 - Центр координат совпадает с мат. ожиданием наблюдений
 - Первый вектор направлен таким образом, что дисперсия вдоль него была максимальной
 - Каждый последующий вектор ортогонален предыдущим и направлен по направлению максимальной дисперсии
 - Последние компоненты – не важны!!!

- Формально:

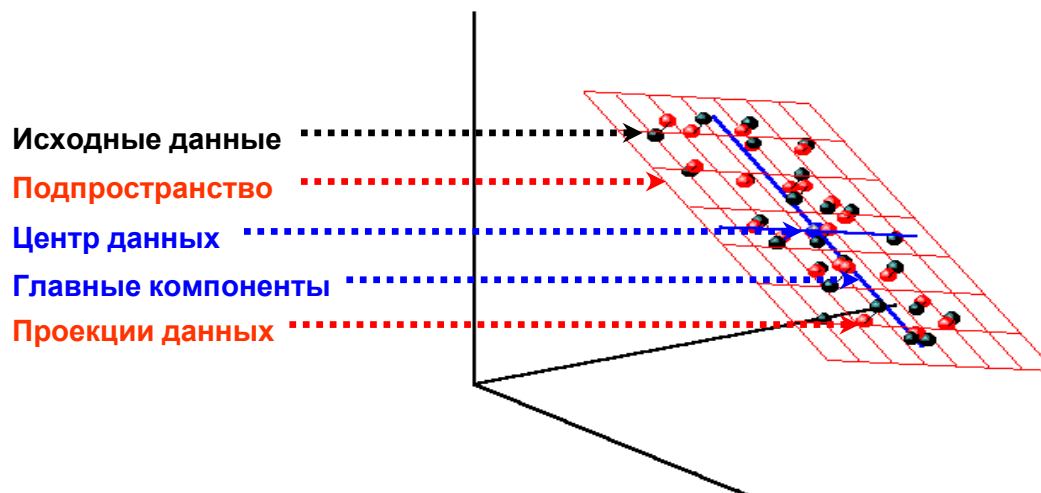
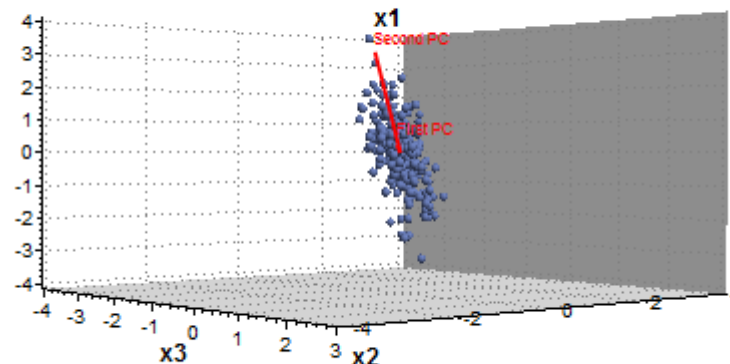
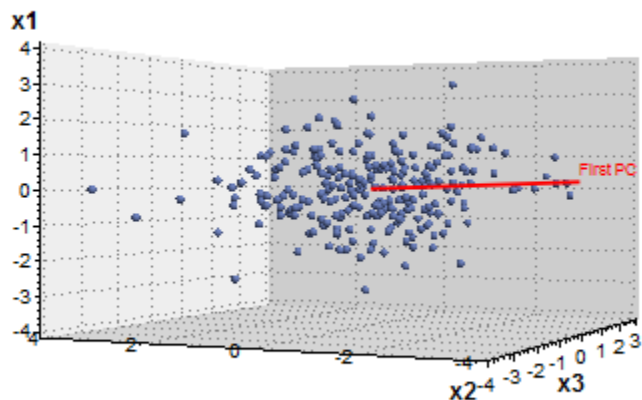
$$\mathbf{w}_1 = \arg \max_{\|\mathbf{w}\|=1} E\{(\mathbf{w}^T \mathbf{x})^2\}$$

$$\mathbf{w}_k = \arg \max_{\|\mathbf{w}\|=1} E\left[\mathbf{w}^T \left(\mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x}\right)\right]^2$$

- - SVD разложение матрицы данных
 - Собственные значения ковариационной матрицы



Геометрическая интерпретация



Поиск собственных значений и собственных векторов ковариационной

- Рассчитаем (или корреляционную) ковариационную матрицу:

- Ковариация = 0 – независимы
- Ковариация > 0 – вместе растут и убывают
- Ковариация < 0 – противофаза

$$C = \begin{bmatrix} cov(x_1, x_1) & \cdots & cov(x_1, x_m) \\ \vdots & \ddots & \vdots \\ cov(x_m, x_1) & \cdots & cov(x_m, x_m) \end{bmatrix}$$

- Проблема с.зн.:

$$C * v = \lambda * v$$

решение: поиск корней

$$cov(x_i, x_j) = E[(x_i - E(x_i))(x_j - E(x_j))]$$

$$|C - \lambda \cdot I| = 0$$

матрица положительно определенная – есть вещественные корни

- Результат:

- λ – дисперсии, с.в. – главные компоненты

$$z_1 = \alpha_{11}x_1 + \cdots + \alpha_{1m}x_m, \dots, z_k = \alpha_{k1}x_1 + \cdots + \alpha_{km}x_m,$$

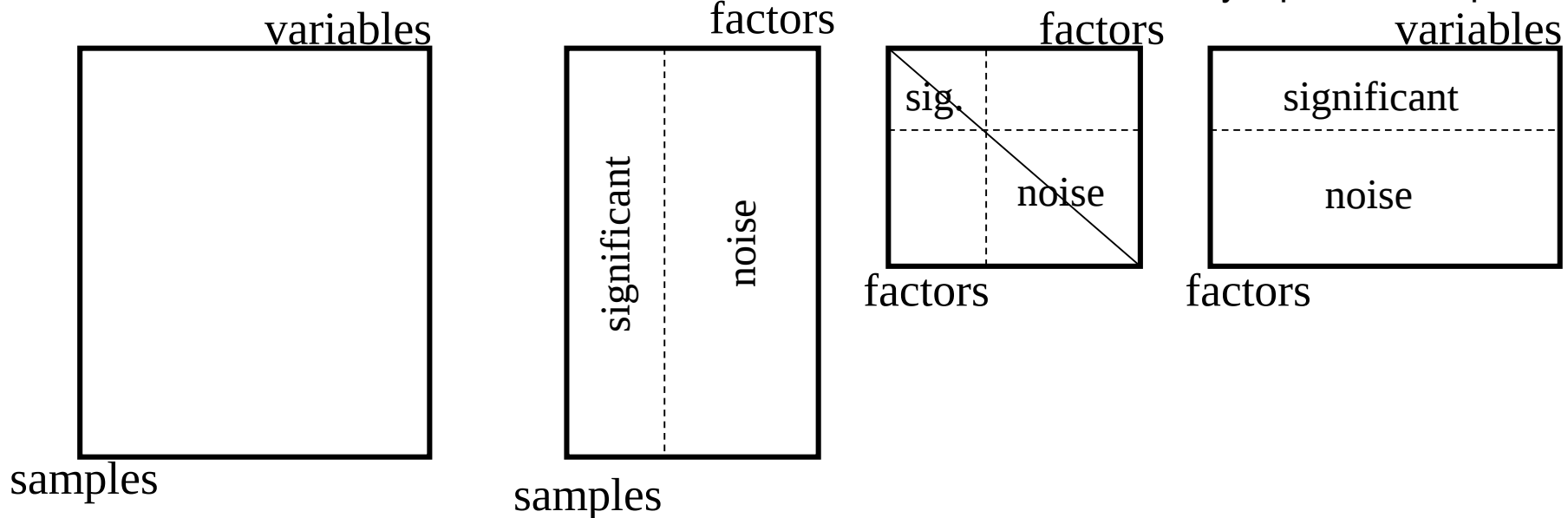
$$\forall i \sum_{j=1}^m \alpha_{ij}^2 = 1$$

Сингулярное разложение в PCA

$$X = U S V^T$$

Унитарная матрица левых
сингулярных векторов

Унитарная матрица правых
сингулярных векторов



■ Если тексты:

- X – матрица корпуса документов, samples – документы, variables – термы, factors – тематики, U – представление документов в пространстве тематик, V – представление тематик в пространстве термов, S – веса тематик

■ Если рекомендации:

- X – матрица частот покупок, samples – клиенты, variables – продукты, factors – предпочтения, U – представление клиентов в пространстве предпочтений, V – представление продуктов в пространстве предпочтений, S – веса предпочтений

SVD разложение и обратная проекция

- SVD разложение матрицы X :
$$X_{n \times m} = U_{n \times n} D_{n \times m} V_{m \times m}^T$$
- SVD приближение (метод главных компонент):
 - отбрасываются с.в., соотв. наименьшим с.з.
 - остается p -я часть главных с.в., которые характеризуют основные зависимости в X
$$\min_{U_p, D_p, V_p} \|X - U_p D_p V_p^T\|$$
 - с их помощью приближается исходная матрица:

$$X^{(l+1)} = V_p V_p^T X^{(l)}$$

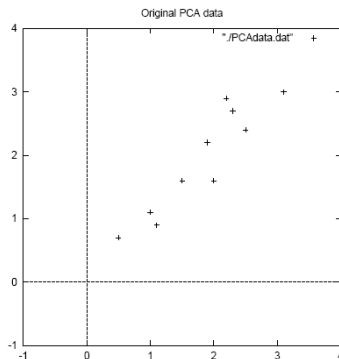


Figure 3.1: PCA example data, original data on the left, data with the means subtracted on the right, and a plot of the data

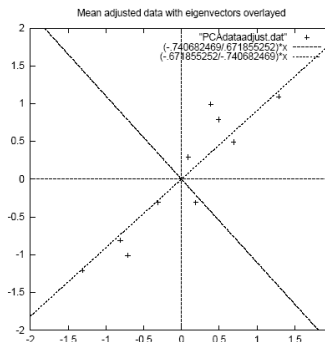


Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlaid on top.

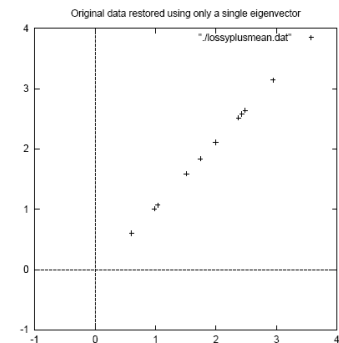


Figure 3.5: The reconstruction from the data that was derived using only a single eigenvector

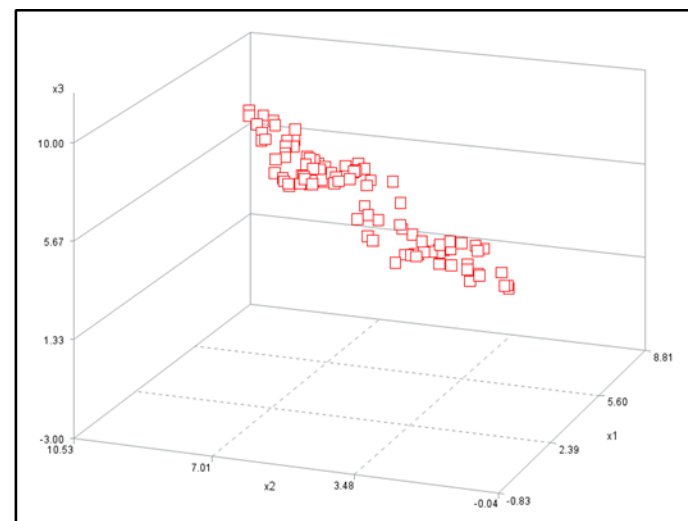
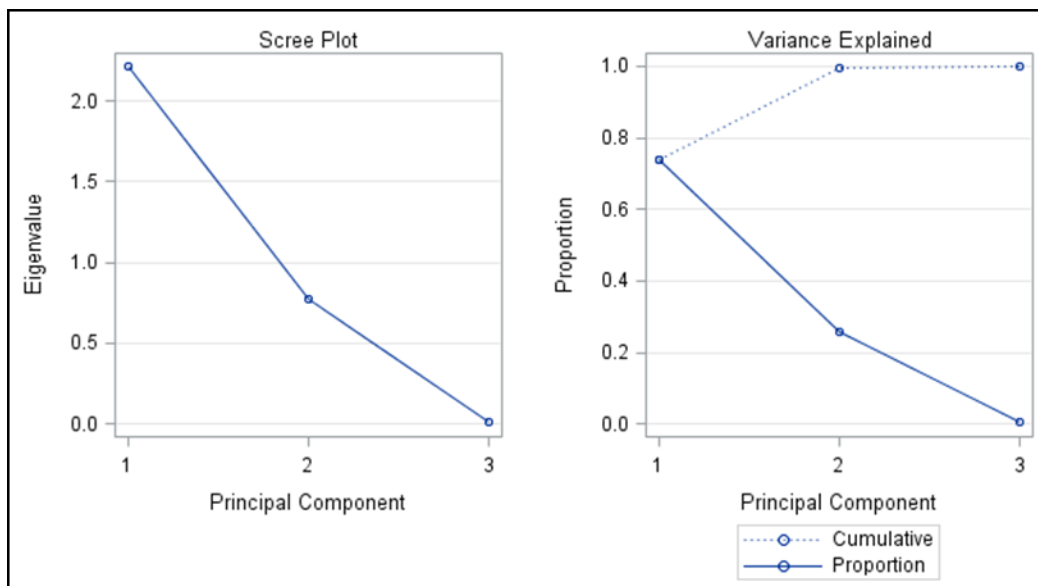
Отбор числа главных компонент

- Пропорция описанной вариации пространства признаков
- По графику
- $S.зн > 1$ (для корреляционной матрицы)
- Теоретические методы:
 - Кайзера (остаются с.в. у которых с.зн. больше среднего с.зн)
 - Правило «сломанной трости» - Набор нормированных на единичную сумму собственных чисел сравнивается с распределением длин обломков трости единичной длины, сломанной в $n-1$ -й случайно выбранной точке (точки разлома выбираются независимо и равномерно распределены по длине трости). Выбираются те компоненты, для которых:

$$\lambda_i > E(L_i) = \frac{1}{n} \sum_{j=i}^n \frac{1}{j}$$

$$\frac{\sum_{i=1}^n z_{im}^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}$$

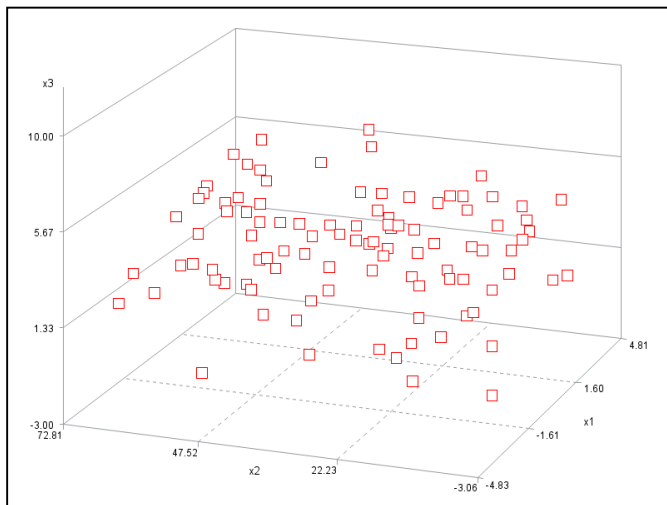
Пример 1



Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.21358154	1.44403082	0.7379	0.7379
2	0.76955072	0.75268297	0.2565	0.9944
3	0.01686775		0.0056	1.0000

Eigenvectors			
	Prin1	Prin2	Prin3
x1	0.650940	0.263685	0.711862
x2	0.645235	0.301851	-.701825
x3	-.399937	0.916164	0.026348

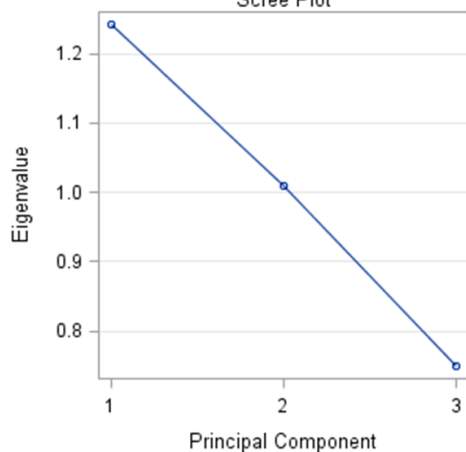
Пример 2



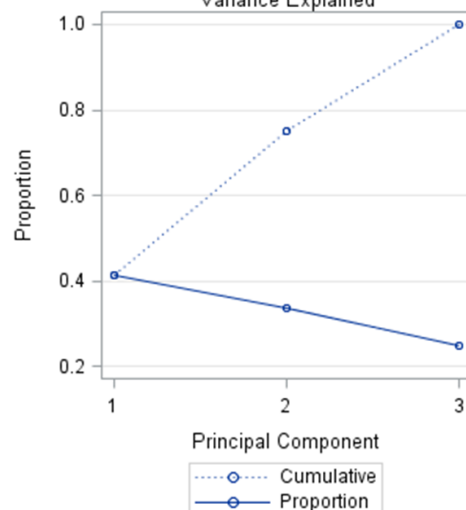
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.24205697	0.23274599	0.4140	0.4140
2	1.00931098	0.26067894	0.3364	0.7505
3	0.74863205		0.2495	1.0000

Scree Plot



Variance Explained

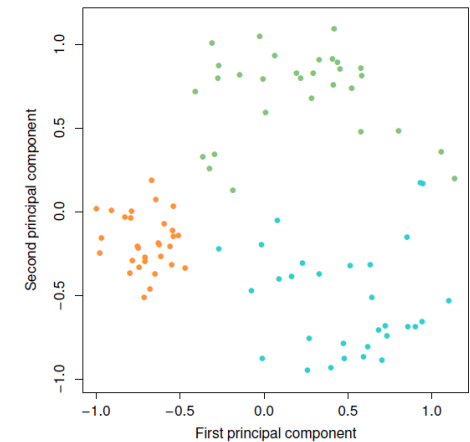
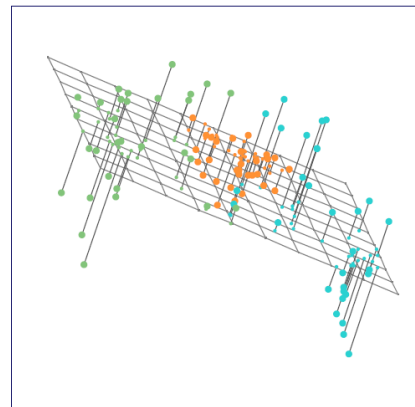
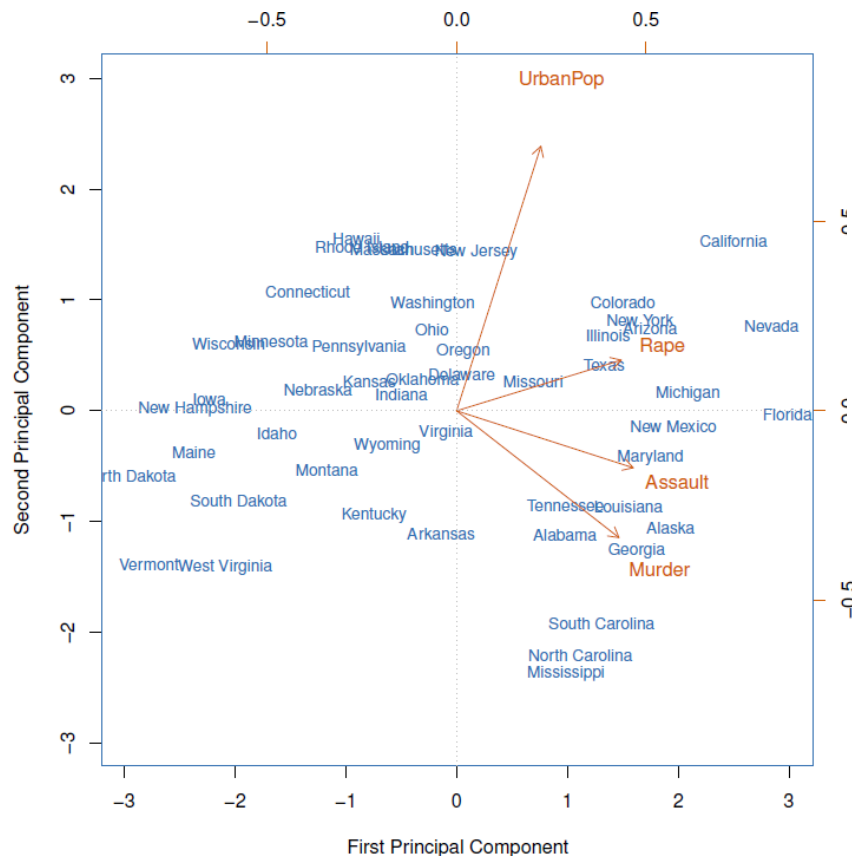


Eigenvectors

	Prin1	Prin2	Prin3
x1	0.704619	-.156546	0.692102
x2	0.708942	0.113759	-.696032
x3	0.030228	0.981097	0.191139

Визуализация

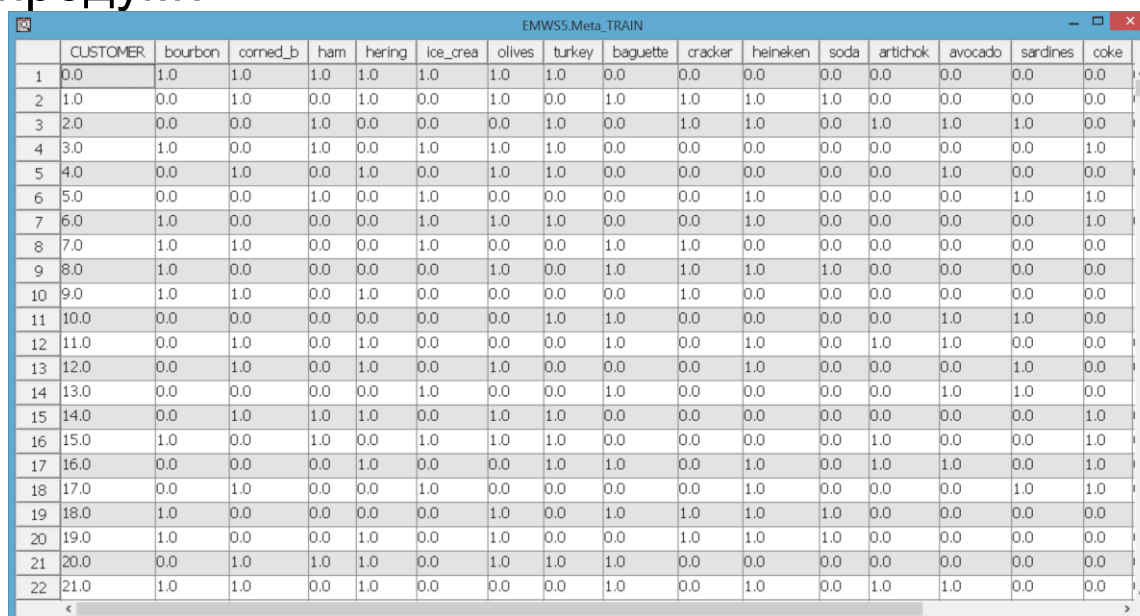
- Одно из важнейших применений – проекция множества наблюдений на пары главных компонент



	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Пример использования в SAS EM

- Пример: преобразуем транзакционный набор ASSOC в матрицу клиент-продукт:

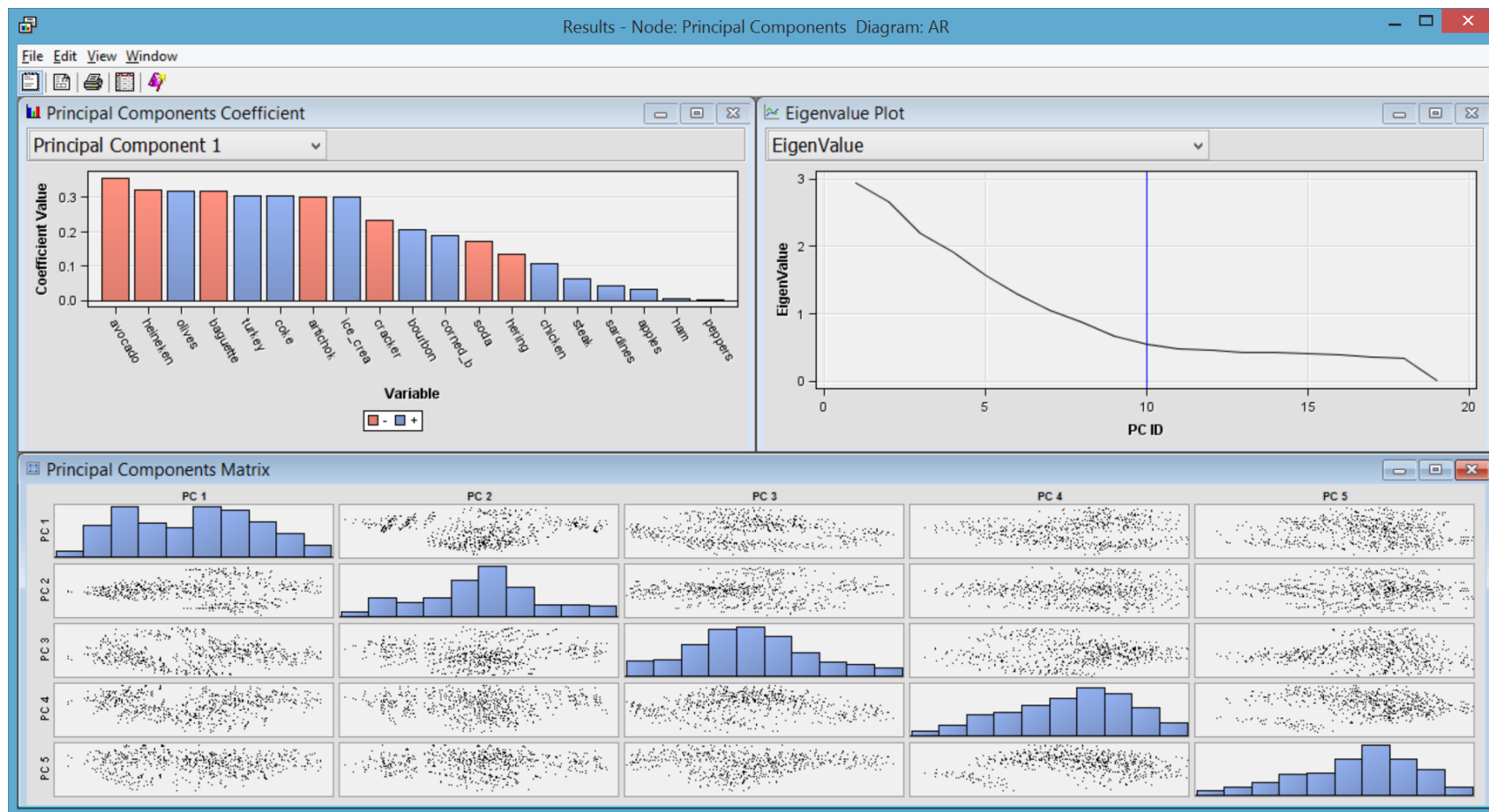


	CUSTOMER	bourbon	corned_b	ham	herring	ice_crea	olives	turkey	baguette	cracker	heineken	soda	artichok	avocado	sardines	coke
1	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
3	2.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0
4	3.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
5	4.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
6	5.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
7	6.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
8	7.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	8.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
10	9.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
11	10.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
12	11.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0
13	12.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
14	13.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
15	14.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
16	15.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
17	16.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0
18	17.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
19	18.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
20	19.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
21	20.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	21.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0

- Применяем метод главных компонент (раздел Modify):



Пример использования в SAS EM



Пример использования в SAS EM

■ Результаты:

Новые признаки клиентов
(матрица U)

EMWSS.PRINCOMP_TRAIN									
	hes	coke	peppers	apples	steak	chicken	Principal Component 1	Principal Component 2	Principal Component 3
1	0.0	0.0	0.0	0.0	0.0	0.0	2.682934990667656	-1.8950185348616713	-0.647252039789035
2	0.0	0.0	0.0	0.0	0.0	0.0	-1.1631019588972604	-1.4609270307611544	-1.635443076641941
3	0.0	0.0	0.0	0.0	0.0	0.0	-1.4916674502952272	0.5903625874052082	0.013127522158427
4	1.0	1.0	0.0	0.0	0.0	0.0	3.2246627397267527	0.4801235590209436	-0.38616817677986
5	0.0	0.0	1.0	1.0	0.0	0.0	1.1182905076795822	-3.139188790794837	1.804750926315555
6	1.0	1.0	0.0	0.0	1.0	1.0	1.1940172691193203	3.114317120293117	0.784083260527779
7	1.0	0.0	1.0	0.0	0.0	0.0	2.6257650798281453	0.9383903971854824	-0.367398354291142
8	0.0	1.0	0.0	0.0	1.0	1.0	0.7630545481708811	1.1043181413115415	-0.080291549175911
9	0.0	1.0	0.0	0.0	0.0	0.0	-0.8759016643139358	0.3309286681402323	-1.986311226797292
10	0.0	1.0	0.0	0.0	1.0	1.0	0.5187547223060062	-0.3935978401697635	-0.462270770362266
11	0.0	1.0	1.0	1.0	0.0	0.0	-0.1897520060401034	-0.11528147557857721	3.488386416254365
12	0.0	0.0	1.0	0.0	0.0	0.0	-2.257445422347318	-1.1165948574243063	1.886624423135866
13	0.0	0.0	1.0	1.0	0.0	0.0	0.6209125163551981	-1.5698048739612693	1.522260029554163
14	0.0	1.0	1.0	1.0	0.0	0.0	-0.22202114343491194	0.9450340027191764	3.636227367049452
15	1.0	0.0	1.0	0.0	0.0	0.0	2.376322919297899	-2.2628021395472344	0.831702932216504
16	1.0	0.0	0.0	0.0	0.0	0.0	2.5857270645773553	0.31045786698337663	-0.77203048776485
17	1.0	0.0	0.0	0.0	0.0	0.0	-1.3901321487340537	0.22030489380084506	0.733931351506303
18	1.0	0.0	1.0	0.0	1.0	1.0	1.632733091055635	2.3977954998494972	1.5188202198011986
19	0.0	1.0	0.0	0.0	0.0	0.0	-0.8759016643139358	0.3309286681402323	-1.986311226797292
20	0.0	0.0	0.0	1.0	0.0	0.0	-0.3495095261304195	-1.2041172590044207	-2.4837622400671697
21	0.0	0.0	1.0	0.0	0.0	0.0	1.0664910707226987	-2.9610071223671084	1.211732540748386
22	0.0	0.0	0.0	0.0	0.0	0.0	-1.9238234097637017	-0.7171673641368124	0.4180913625111569

Новые признаки продуктов
(матрица V)

EMWSS.PRINCOMP_EIGENVECTOR						
Variable	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
1 apples	0.033538839282211956	-0.11561098283925879	0.4232060808928141	-0.20649732018295594	0.17674264634555992	-0.027648377659581374
2 artichok	-0.3022561674280135	0.005733870963486466	0.02984196386866408	0.31317082567104404	-0.273384414331428	-0.05346588067524221
3 avocado	-0.3578922220742492	0.007011003779861447	0.2768864144044813	0.155640982572533	-0.14750548853094797	-0.29516954723768557
4 beguette	-0.3180850161154238	0.020529174371872683	0.20073321467029595	-0.07196675814337779	0.3501790059121384	-0.1735343918691918
5 bourbon	0.20615011441116218	0.08009768615194164	-0.2964653532898179	-0.27805574841281405	-0.07494005261024755	-0.3533758719229611
6 chicken	0.10862278827195285	0.29445574847422806	-0.015751137000675198	-0.17999402180013224	-0.273844538703087	0.4565575598537217
7 coke	0.3041091298237175	0.34186439410567443	0.014321677781409846	0.21665584152930778	0.13597997137711876	0.05628175938184295
8 corned_b	0.18859993927842958	-0.34617280209010026	0.08280850641909114	-0.17566887838855128	-0.2973631607225058	0.27677187927911157
9 cracker	-0.2322125539749572	0.0280030552600619	-0.40544904393589287	-0.26369501190856404	-0.1794382715396897	0.023577659862706324
10 ham	0.005843450396472283	-0.19258540615980355	-0.059078275779843	0.34648778516001116	-0.43668452066262775	-0.04021666134587923
11 heineken	-0.3225440406210939	0.23159253938743954	-0.27022773432513697	0.20213702528972766	0.08300507079148092	0.21245891178312537
12 hering	-0.1331572457936531	-0.3766868184510856	0.02502755184766046	0.12796701804862612	0.25510691595391594	0.26317995755941
13 ice_crea	0.3008506912047889	0.3375912891344994	0.009809016153991526	0.22251359967967704	0.11627247468883248	0.025287007408597933
14 olives	0.31827207982535	-0.29522625255275087	-0.17215797575866712	0.05698322479755659	0.2006541594143905	-0.1364095260442434
15 peppers	0.0017384905145768675	0.08625006346324338	0.12191536021216093	-0.4733862404131001	-0.2342114904873427	-0.27243483348283937
16 sardines	0.04134836306076211	0.35241102489044657	0.32421789737057766	-0.005392826343645234	0.09152653247827508	0.06860690758394204
17 soda	-0.17295687499859633	-0.015039040531984684	-0.40386230560203684	-0.1896127796735675	0.3636447772281753	-0.019318319732229396
18 steak	0.06403676491251577	-0.252792420725851	0.14007393297216267	-0.0682570858078998	0.13363247854653565	0.3605830461790619
19 turkey	0.30437522360739716	-0.161286532628298	-0.05835384211206471	0.2793717683844141	0.012166871637040132	-0.34771304907918865

PC ID	Principal components	EigenValue
1 1.0	PC_1	2.941055136636557
2 2.0	PC_2	2.6532280980422387
3 3.0	PC_3	2.2012673929073814
4 4.0	PC_4	1.9122988981046196
5 5.0	PC_5	1.5626739830550092
6 6.0	PC_6	1.2857819883561596
7 7.0	PC_7	1.045845752505471
8 8.0	PC_8	0.8741575178125174
9 9.0	PC_9	0.6708700141827175

Веса предпочтений (синг. числа lambda) =>

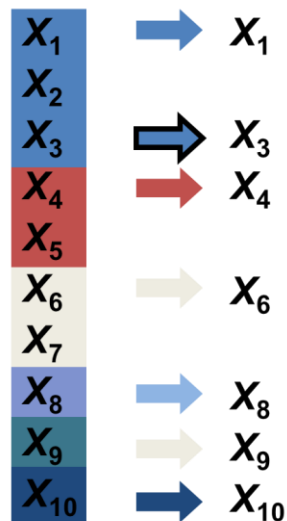
■ Как узнать купит ли клиент u продукт v?

- Посчитать скалярное произведение $\langle u, v \rangle * \lambda$

Кластеризация переменных

■ Задачи:

- группировка переменных в иерархические кластеры так, чтобы в одном кластере переменные были **максимально коррелированы**, а кластеры между собой нет
- Затем выбирается либо первая гл. компонента кластера либо лучший представитель кластера



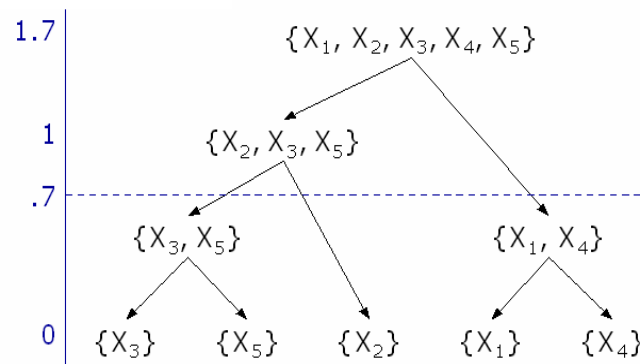
Выбор переменных

- Лучший представитель
- Экспертный выбор
- Корреляция с откликом

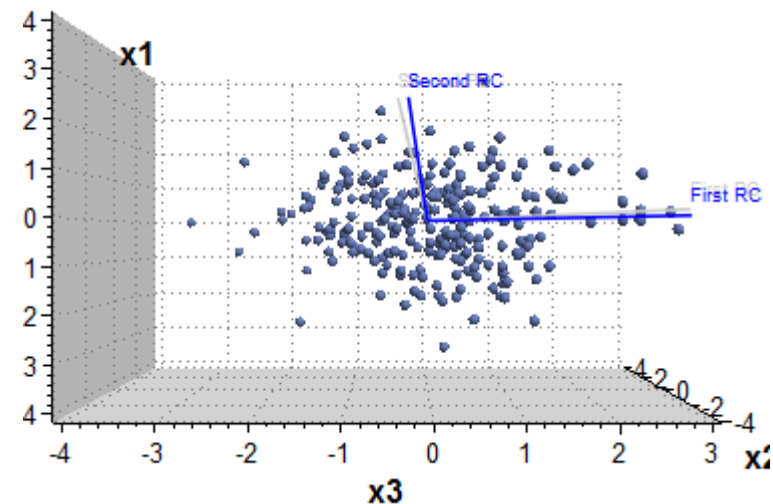
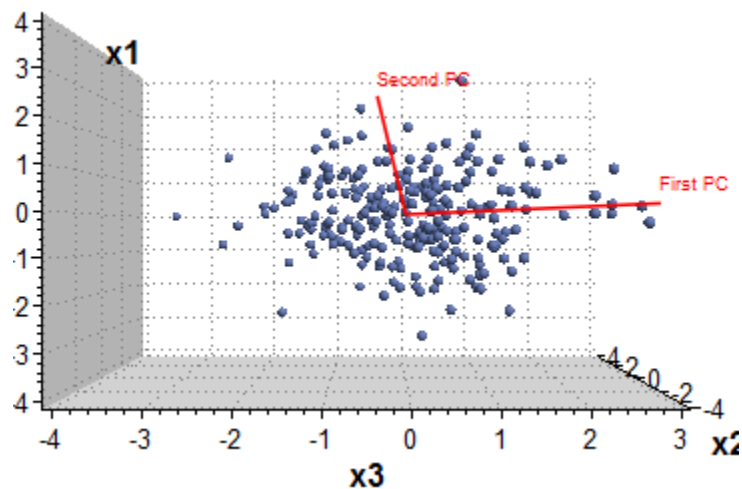
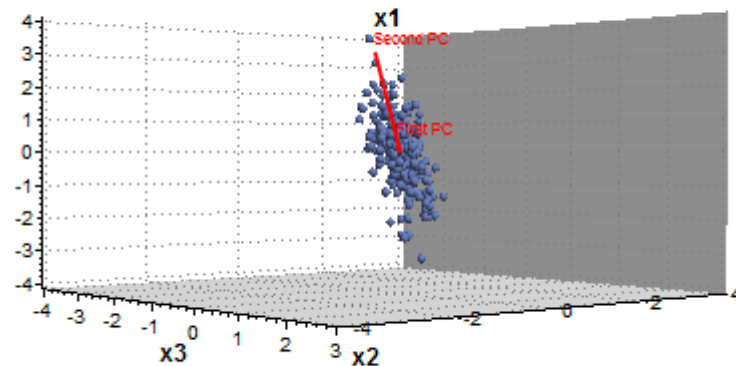
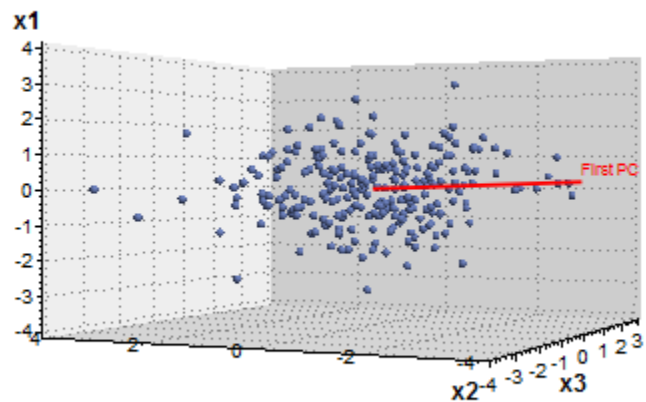
Суть алгоритма группировки переменных

- Нисходящая иерархическая кластеризация, сначала все переменные в одном кластере
- Затем повторяется процесс:
 1. Выбор кластера (группы коррелирующих переменных) для разбиения
 2. Деление кластера на два с помощью метода гл. комп. с вращением
 3. Перераспределение переменных по кластерам

2nd Eigenvalue



Шаг разбиения



Поворот главных компонент

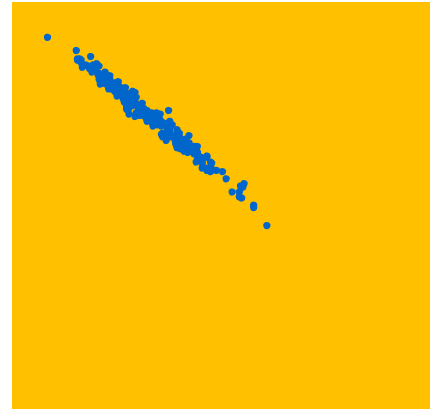
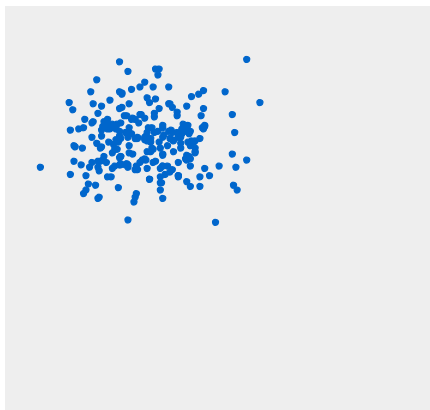
Перераспределение переменных

X_1

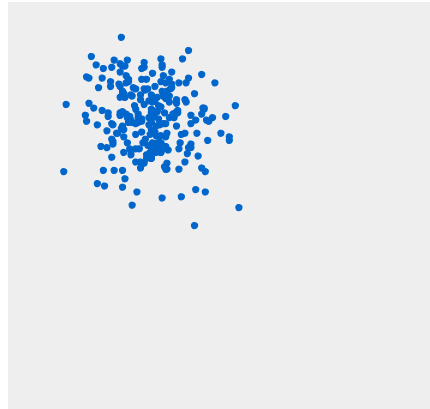
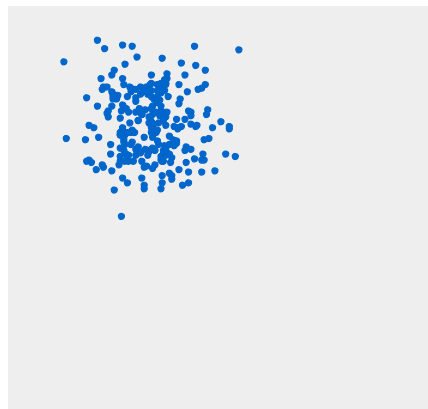
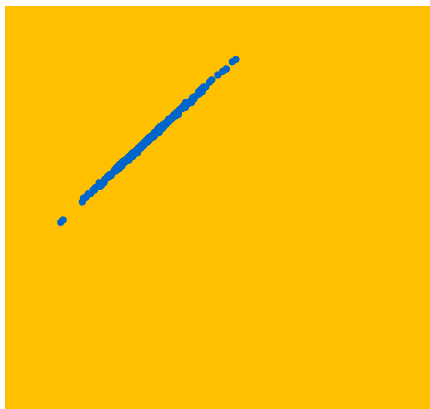
X_2

X_3

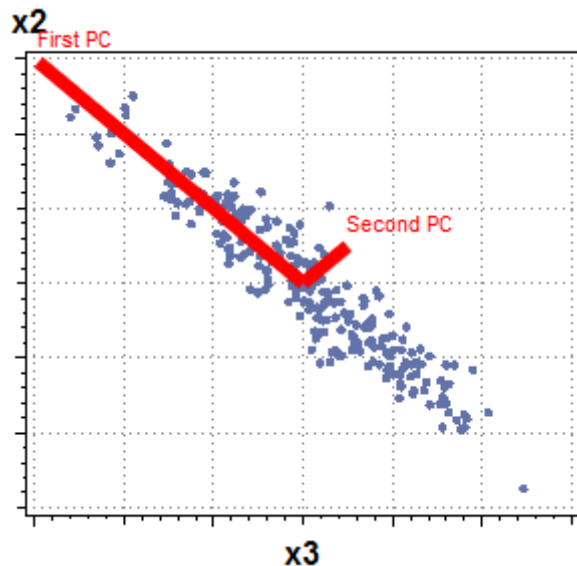
First
RC



Second
RC

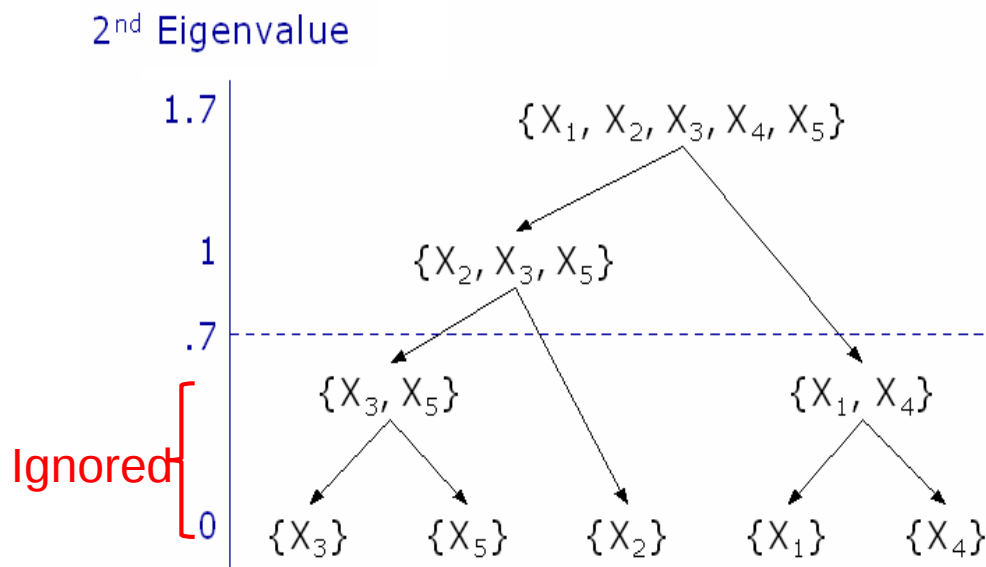


Разбивать ли дальше?



First Eigenvalue=1.95

Second Eigenvalue=0.05

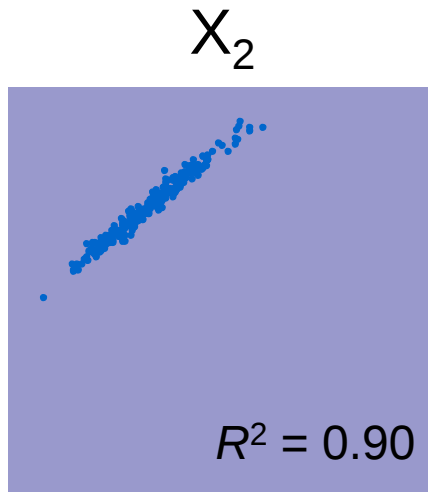


Варианты остановки (помимо порога на с.зн.):

- Задать максимальное число кластеров
- Задать минимум описанной вариации (дисперсии)

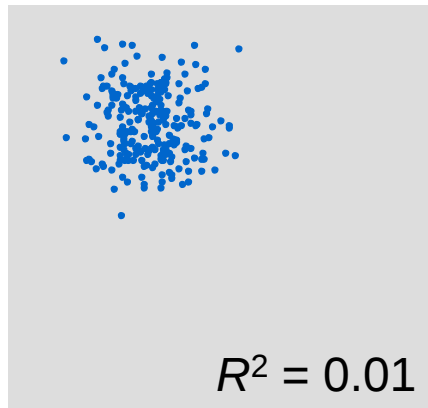
Выбор представителей кластеров

First
Cluster
PC



$$\frac{1 - R^2_{own\ cluster}}{1 - R^2_{next\ closest}} = \frac{1 - 0.90}{1 - 0.01} = 0.101$$

Second
Cluster
PC

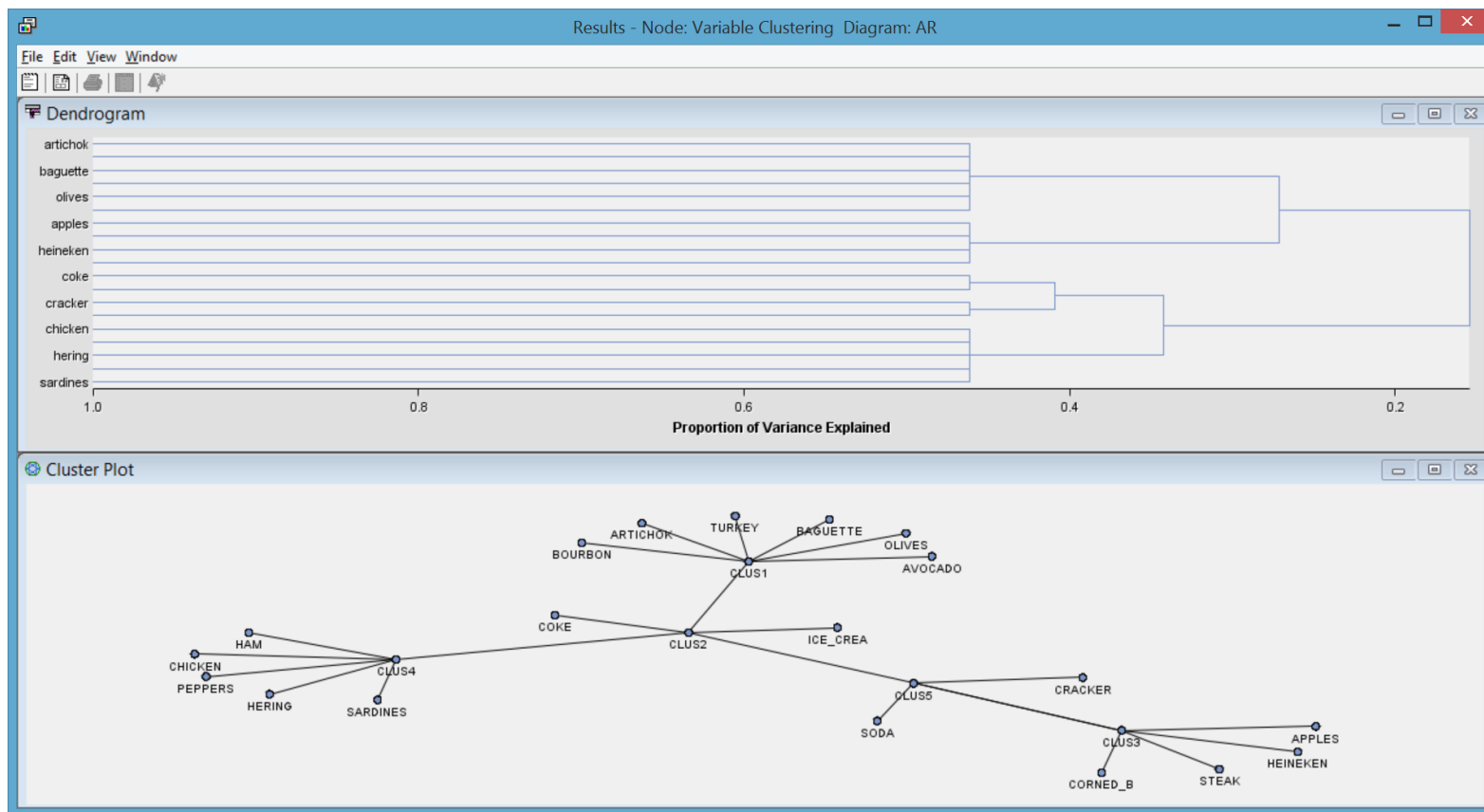


Также можно выбирать не
автоматически:

- Эксперт «вручную»
- По корреляции с откликом

Пример использования в SAS EM

- В разделе Explore узел:



Пример использования в SAS EM

■ Отбор переменных:

Variable Selection Table								
Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	Label	1-R2 Ratio	Variable Selected
CLUS1	CLUS1	1	CLUS2	0.078861	ClusterComp	Cluster 1	0	NO
CLUS1	AVOCADO	0.542912	CLUS2	0.093912	Variable		0.504463	YES
CLUS1	OLIVES	0.521717	CLUS4	0.110853	Variable		0.537912	NO
CLUS1	BAGUETTE	0.325976	CLUS2	0.059756	Variable		0.716861	NO
CLUS1	ARTICHOK	0.331882	CLUS3	0.073638	Variable		0.721228	NO
CLUS1	BOURBON	0.277107	CLUS4	0.036053	Variable		0.74993	NO
CLUS1	TURKEY	0.269857	CLUS5	0.076499	Variable		0.790625	NO
CLUS2	CLUS2	1	CLUS5	0.104983	ClusterComp	Cluster 2	0	NO
CLUS2	ICE_CREA	0.784811	CLUS5	0.080811	Variable		0.234108	YES
CLUS2	COKE	0.784811	CLUS5	0.083987	Variable		0.234919	NO
CLUS3	CLUS3	1	CLUS5	0.133438	ClusterComp	Cluster 3	0	NO
CLUS3	HEINEKEN	0.652463	CLUS1	0.122696	Variable		0.396142	YES
CLUS3	CORNEB	0.463746	CLUS1	0.075951	Variable		0.580331	NO
CLUS3	APPLES	0.360347	CLUS5	0.076063	Variable		0.692313	NO
CLUS3	STEAK	0.298494	CLUS4	0.028484	Variable		0.722073	NO
CLUS4	CLUS4	1	CLUS2	0.096784	ClusterComp	Cluster 4	0	NO
CLUS4	HERING	0.471329	CLUS2	0.135546	Variable		0.611567	YES
CLUS4	SARDINES	0.402806	CLUS2	0.09078	Variable		0.65682	NO
CLUS4	CHICKEN	0.339126	CLUS2	0.048454	Variable		0.694526	NO
CLUS4	PEPPERS	0.328812	CLUS3	0.041087	Variable		0.699947	NO
CLUS4	HAM	0.205224	CLUS2	0.018035	Variable		0.809373	NO
CLUS5	CLUS5	1	CLUS3	0.133438	ClusterComp	Cluster 5	0	NO
CLUS5	SODA	0.70086	CLUS3	0.097907	Variable		0.331606	YES
CLUS5	CRACKER	0.70086	CLUS2	0.106259	Variable		0.334705	NO

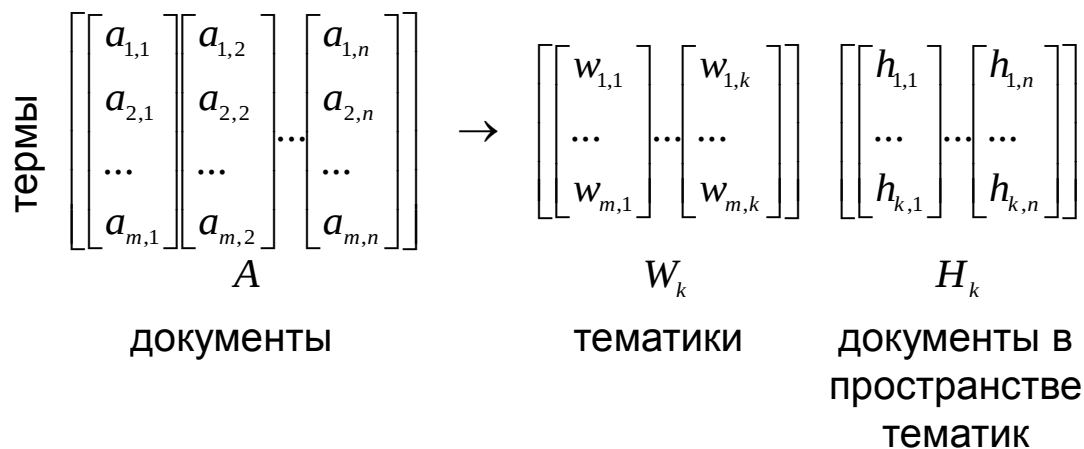
Дополнительно
новые координаты

EMWS5.VarClus_TRAIN				
apples	steak	chicken	Clus1	Clus2
0.0	0.0	0.0	-1.5643885061949059	0.4444970068771528
0.0	0.0	0.0	-0.09554399287678009	-0.7401707103423649
0.0	0.0	0.0	0.7152518121698608	-0.7401707103423649
0.0	0.0	0.0	-1.5643885061949059	1.6665995135111098
1.0	1.0	0.0	-0.44774227582526505	-0.7401707103423649
0.0	0.0	1.0	0.018769217840735486	1.6665995135111098
1.0	0.0	0.0	-1.5643885061949059	1.6665995135111098
0.0	0.0	1.0	0.07939532206166983	0.4444970068771528
0.0	0.0	0.0	-0.5500723017920117	-0.7401707103423649
0.0	0.0	1.0	-0.4357590910744962	-0.7401707103423649
1.0	1.0	0.0	0.6968797611645825	-0.7401707103423649
1.0	0.0	0.0	1.7295680165727552	-0.7401707103423649
1.0	1.0	0.0	-0.610698406012946	-0.7401707103423649
1.0	1.0	0.0	1.1960415524313106	0.4444970068771528
1.0	0.0	0.0	-1.1098601972796742	0.48193179629159194
0.0	0.0	0.0	-1.0308620420534615	1.6665995135111098
0.0	0.0	0.0	1.2304062253060268	0.48193179629159194
1.0	0.0	1.0	0.018769217840735486	1.6665995135111098
0.0	0.0	0.0	-0.5500723017920117	-0.7401707103423649
0.0	1.0	0.0	1.0652267140221776	0.7401707103423649

Неотрицательная матричная факторизация

- $A \in \mathbb{R}^{m \times n}$, $(a_{ij} \geq 0)$
- Цель неотрицательной матричной факторизации состоит в нахождении матриц $W_k \in \mathbb{R}^{m \times k}$ и $H_k \in \mathbb{R}^{k \times n}$ с неотрицательными элементами, которые минимизируют целевую функцию:

$$f(W_k, H_k) = \frac{1}{2} \|A - W_k H_k\|_F^2, k \ll \min(m, n)$$



Ортонормированная неотрицательная матричная факторизация (ОНМФ)

Целевая функция: $f(W_k, H_k) = \frac{1}{2} \|A - W_k H_k\|_F^2 + \frac{\alpha}{2} \|W_k^T W_k - I\|_F^2$

1. $W^1 \in \mathbb{R}^{m \times k}$, $H^1 \in \mathbb{R}^{k \times n}$ инициализируются случайными неотрицательными числами
2. В цикле p раз выполняются итерационные формулы для вычисления матриц W и H :

$$H_{b,j}^{p+1} = H_{b,j}^p \frac{((W^p)^T A)_{b,j}}{((W^p)^T W^p H^p)_{b,j}}, \forall b, j: 1 \leq b \leq k, 1 \leq j \leq n$$

$$W_{i,a}^{p+1} = W_{i,a}^p \frac{(A(H^{p+1})^T + \alpha W^p)_{i,a}}{(W^p H^{p+1} (H^{p+1})^T + \alpha W^p (W^p)^T W^p)_{i,a}}, \forall i, a: 1 \leq i \leq m, 1 \leq a \leq k$$