# Лекция 11

## Методы кластерного анализа проектирование данных на плоскость, метод главных компонент

**Лектор** – ***Сенько Олег Валентинович***

Курс «Математические основы теории прогнозирования»
4-й курс, III поток

# Содержание лекции

Важное прикладное значение имеют методы анализа данных, связанные с теорией распознавания. К их числу относятся методы кластерного анализа и методы визуализации многомерных данных. Целью методов кластерного анализа является разбиение выборок многомерных данных на группы объектов близких в смысле некоторой заданной меры сходства. Такие компактные группы называются кластерами, классами или таксонами. Методы кластерного анализа называют также методами обучения без учителя, автоматической группировки или таксономии. Методы кластерного анализа могут использоваться в качества вспомогательных инструментов при решении задач прогнозирования или распознавания. Так с помощью кластеризации могут отбираться эталонные объекты. Однако нередко кластеризация может иметь самостоятельное значение. Можно выделить задачи кластерного анализа, для которых число кластеров задано, а также задачи, в которых число кластеров следует определить в ходе решения кластеризации.

Большинство известных алгоритмов кластеризации предполагает задание неотрицательной функцией близости $\rho(\boldsymbol{x}, \boldsymbol{y})$ между произвольными векторами $\boldsymbol{x}$ и $\boldsymbol{y}$. В качестве функций близости могут выступать евклидова метрика или метрика Хэмминга. Одним из наиболее известных методов кластеризации является алгоритм k внутригрупповых средних. Предположим, что у нас задана выборка многомерных векторов-объектов $\widetilde{S}_{ini} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$. Алгоритм находит такие кластеры, для объектов которых центр «своего кластера» будет ближе центра любого «чужого кластера».На начальном этапе произвольным образом выбирается начальная кластеризация $\widetilde{G}_0 = \{G_1^0, \dots, G_k^0\}$ с содержанием объектов $(m_1^0 \dots, m_k^0)$ соответственно.
Предположим, что на шаге $(l-1)$ получены группы $\{G_1^{l-1}, \dots, G_k^{l-1}\}$ с содержанием объектов $(m_1^{l-1}, \dots, m_k^{l-1})$ На шаге $l$ для каждой из групп $G_i^{l-1}$ вычисляется центр

$$\bar{\boldsymbol{x}}_i^{l-1} = \frac{1}{m_i^l} \sum_{\boldsymbol{x}_j \in G_i^{l-1}} \boldsymbol{x}_j$$

при $i = 1, \dots, k$.

Произвольный объект $\boldsymbol{x}_{j'}$ из выборки $\widetilde{S}_i ni$ переносится в группу $G_{i''}^{l-1}$ , если при произвольном $i'$ из множества $\{1,\ldots,k\} \setminus i''$ выполняется неравенство

$$\rho(\boldsymbol{x}_{\boldsymbol{j'}}, \bar{\boldsymbol{x}}_{i''}^{l-1}) < \rho(\boldsymbol{x}_{\boldsymbol{j'}}, \bar{\boldsymbol{x}}_{i'}^{l-1}).$$

. В результате мы получаем группы $\{G_1^l,\ldots,G_k^l\}$ и переходим к шагу $(l+1)$. Процесс останавливается, если на каком-то шаге оказывается, что $\bar{\boldsymbol{x}}_i^l = \bar{\boldsymbol{x}}_i^{l+1}$ при $i = 1,\ldots,k$.

Другим методом кластеризации, основанным на итерационной процедуре является алгоритм Форель, основанный на. движении гипершаров фиксированного радиуса в сторону мест «сгущения»объектов. Пусть фиксировано некоторое положительное число $R$. Выбирается случайный вектор $\boldsymbol{x}_{j'} \in \widetilde{S}_{ini}$ и гипершар $\boldsymbol{R}_1$ радиуса $R$ с центром в $\boldsymbol{z}_1 = \boldsymbol{x}_{j'}$. То есть $\boldsymbol{R}_1 = \{\boldsymbol{x} \mid \rho(\boldsymbol{x}, \boldsymbol{z}_1) < R\}$. Полагаем $G_1 = \widetilde{S}_{ini} \bigcap \boldsymbol{R}_1$ и вычисляем центр новой сферы по формуле

$$\boldsymbol{z}_2 = \frac{1}{\mid G_1 \mid} \sum_{\boldsymbol{x}_j \in G_1} \boldsymbol{x}_j.$$

Формируем группу $G_2$ из объектов $\widetilde{S}_{ini}$, попавших в сферу $\boldsymbol{R}_2 = \{\boldsymbol{x} \mid \rho(\boldsymbol{x}, \boldsymbol{z}_2) < R\}$. Процесс заканчивается на некотором шаге $l^*$ при выполнении условия $G_{l^*+1} = G_{l^*}$. . Полученное множество объектов объявляется первым кластером $G_1^f$ . Оно исключается из $\widetilde{S}_{ini}$ , а вышеописанная процедура повторяется относительно оставшейся части выборки, в результе чего формируется не пересекающаяся с $G_1^f$ выборка $G_2^f$

Процесс кластеризации заканчивается на итерации $k^*$, на которой достигается условие

$$\widetilde{S}_{ini} \setminus \cup_{i=1}^{k} G_i^f = \emptyset.$$

Таким образом формируется набор кластеров

$$G_1^f, \ldots, G_{k^*}^f.$$

Полученное число кластеров зависит от выбора радиуса $R$, который является параметром алгоритма.

Метод иерархической группировки позволяет не только осуществить кластеризацию с заранее выбранным числом классов и выявить иерархию кластеров. На начальном этапе в качестве кластеров рассматриваются отдельные объекты выборки $\widetilde{S}_{ini}$. Дальнейшая кластеризация производится с используется функции близости между кластерами, которая задаётся на основе функции близости между векторными описаниями объектов. На практике используется несколько типов функций близости между кластерами $G_{i'}$ и $G_{i''}$:

- минимальное расстояние между объектами из двух кластеров

$$\mathrm{P}_{min}(G_{i'}, G_{i''}) = \min_{\boldsymbol{x}_\mu \in G_{i'}, \boldsymbol{x}_\nu \in G_{i''}} \rho(\boldsymbol{x}_\mu, \boldsymbol{x}_\nu);$$

- максимальное расстояние между объектами из двух кластеров

$$\mathrm{P}_{max}(G_{i'}, G_{i''}) = \max_{\boldsymbol{x}_\mu \in G_{i'}, \boldsymbol{x}_\nu \in G_{i''}} \rho(\boldsymbol{x}_\mu, \boldsymbol{x}_\nu);$$

- расстояние между центрами двух кластеров

$$\mathrm{P}_c(G_{i'}, G_{i''}) = \rho(\bar{\boldsymbol{x}}_{i'}, \bar{\boldsymbol{x}}_{i''});$$

- среднее расстояние между объектами двух классов

$$\mathrm{P}_{av}(G_{i'}, G_{i''}) = \frac{1}{|G_{i'}|}\frac{1}{|G_{i''}|}\sum_{\boldsymbol{x}_\mu \in G_{i'}}\sum_{\boldsymbol{x}_\mu \in G_{i'}} \rho(\boldsymbol{x}_\mu, \boldsymbol{x}_\nu).$$

На втором шаге два ближайших кластера объединяются в один. Процесс объединения повторяется до выделения заранее фиксированного числа кластеров. Для остановки процесса объединения кластеров могут быть использованы дополнительные условия, задаваемые экспертом, и связанные со спецификой конкретной задачи. В этом случае число кластеров устанавливается в ходе решения.

Используются также методы кластеризации, основанные на поиске разбиений $\widetilde{S}_{ini}$ , для которых достигают максимума специальные функционалы качества.

Так качество разбиения на набор кластеров $\widetilde{G} = \{G_1, \ldots, G_k\}$ может быть описано с помощью функционала внутренних дисперсий $F_{VS}(\widetilde{G})$ представляющего собой взвешенную сумму средних отклонений от центра внутри внутри каждой из групп

$$F_{VS}(\widetilde{G}) = \sum_{i=1}^{k} |G_i| \sum_{\boldsymbol{x}_\mu \in G_i} \frac{\rho(\boldsymbol{x}_\mu, \bar{\boldsymbol{x}}_i)}{|G_i|} = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_\mu \in G_i} \rho(\boldsymbol{x}_\mu, \bar{\boldsymbol{x}}_i).$$

Нетрудно видеть, что "вес" каждой из групп пропорционален числу объектов в ней. Поскольку число всевозможных разбиений $\widetilde{S}_{ini}$ на $k$ групп оценивается как $\frac{k^m}{k!}$ полный перебор разбиений здесь заведомо исключен. Поэтому обычно применяют методы частичного перебора с использованием случайного выбора начальных разбиений и последующей локальной оптимизацией

В методах локальной оптимизации (для определенности, минимизации) строится последовательность разбиений

$$\widetilde{G}_1, \widetilde{G}_2, \ldots, \widetilde{G}_l, \ldots$$

, для которых

$$F_{VS}(\widetilde{G}_1) > F_{VS}(\widetilde{G}_2), \ldots, F_{VS}(\widetilde{G}_l) > F_{VS}(\widetilde{G}_{l+1}), \ldots$$

а разбиение $\widetilde{G}_{l+1}$ вычисляется непосредственно по предшествующему разбиению $\widetilde{G}_l = \{G_1^l, \ldots, G_k^l\}$ путем его «локального» изменения – переноса некоторого объектов из одного кластера в другой. Ищется такой объект $\boldsymbol{x}_{J(l)}$ , при переносе которого из кластера $G_\mu^l$, содержащего $\boldsymbol{x}_{J(l)}$ в разбиении $\widetilde{G}_l$ , в некоторый кластер $G_\nu^l$ уменьшение функционала $F_{VS}$ максимально среди всевозможных переносов такого рода. В результате разбиение отличается от только составом кластеров с номерами $G_\mu^l$ и $G_\nu^l$. Процесс завершается, когда никакой последующий перенос не уменьшает функционал или достигнуто указанное пользователем максимальное число итераций

# Решение задачи кластерного анализа коллективами алгоритмов

Пусть в результате применения разнообразных методов кластеризации получено множество различных решений для одних и тех же данных. При отсутствии внешнего критерия, выбор одного решения из данного множества кластеризаций может быть не ясен. Поэтому представляет интерес применение методов обработки полученных множеств кластеризаций с целью построения коллективных решений, более предпочтительных и обоснованных, чем полученные отдельными алгоритмами кластеризации. Кластеризацию выборки $\widetilde{S}_{ini}$, включающую кластеры $\{G_1, \dots, G_k\}$ можно описать с помощью информационной матрицы $\|\alpha_{ji}\|_{m\times k}$ , где $\alpha_{ji} = 1$, если $\boldsymbol{x}_j \in G_i$, и $\alpha_{ji} = 0$ в противном случае. Наличие нескольких единиц в одной строке соответствует принадлежности объекта сразу нескольким кластерам. Нулевая строка означает отказ от кластеризации соответствующего объекта.

**Определение 1.** Информационные матрицы $I = \|\alpha_{ji}\|_{m\times k}$ и $I = \|\alpha'_{ji}\|_{m\times k}$ называются эквивалентными, если они равны с точностью до перестановки столбцов.

Таким образом произвольная информационная матрица $I = \|\alpha_{ji}\|_{m\times k}$ определяет класс всех эквивалентных ей матриц $\widetilde{K}(I)$.

**Определение 2.** Алгоритмом кластеризации $A^c$ называется алгоритм, переводящий выборку в класс эквивалентности $\widetilde{K}(I)$ некоторой информационной матрицы $I$.

Иными словами $A^c = \widetilde{K}(\|\alpha_{ji}\|_{m\times k})$. Данное определение отражает возможности произвола в обозначении полученных алгоритмом кластеров. Пусть существует $r$ кластеризаций выборки $\widetilde{S}_{ini}$ алгоритмами $A^c_1, \ldots, A^c_r$ на $k$ кластеров. Задача построения оптимальной коллективной кластеризации состоит в вычислении по множеству из $r$ исходных кластеризаций, задающих классы эквивалентности

$$\widetilde{K}(\|\alpha^1_{ji}\|_{m\times k}), \ldots, \widetilde{K}(\|\alpha^r_{ji}\|_{m\times k})$$

некоторого нового коллективного решения $\widetilde{K}(\|\hat{\alpha}_{ji}\|_{m\times k})$ , где $\hat{\alpha}_{ji} \in [0, 1]$.

Оператор $\mathbf{B}(I_1,\dots,I_r) = \|\beta_{ji}\|_{m\times k}$ , где $\beta_{ji} \in \{1,\dots,\}$ , называется сумматором, если

$$\beta_{ji} = \sum_{t=1}^{r} \alpha_{ji}^t.$$

Матрицу, полученную в результате применения сумматора к некоторому набору информационных матриц, будем называть матрицей оценок. Оператор $\mathbf{C}$ называется решающим правилом, если

$$\mathbf{C}(\|\beta_{ji}\|_{m\times k}) = \|\alpha_{ji}^s\|_{m\times k},$$

где при произвольном $j \in \{1,\dots,m\}$ $\alpha_{ji}^s = 1$, если $\beta_{ji} > \beta_{jt}$ при $t = \{1,\dots,k\} \setminus \{i\}$, и $\alpha_{ji}^s = 0$ в противном случае.

**Определение 3.** Комитетным синтезом информационной матрицы $\|\alpha_{ji}^s\|_{m\times k}$ по множеству исходных кластеризаций, задаваемых набором инфориационными матриц $\widetilde{I} = \{I_1,\dots,I_r\}$, называется последовательное применение к $\widetilde{I}$ сумматора $\mathbf{B}$ и решающего правила $\mathbf{C}$.

Для оценивания коллективного решения вводится понятие контрастных матриц оценок, соответствующих случаям, когда все исходные решения задач классификации оказались одинаковыми. Очевидно, что для произвольной контрастной матрицы $\|\beta^c_{ji}\|_{m\times k}$ выполняется условие $\beta^c_{ji} \in \{0, r\}$. Пусть $\widetilde{\mathbf{B}}^c = \|\beta^c_{ji}\|_{m\times k}$ - множество всевозможных контрастных матриц . Качеством произвольной матрицы $B = \|\beta_{ji}\|_{m\times k}$, вычисленной сумматором, определяется как минимальное расстояние до матриц из множества $\widetilde{\mathbf{B}}^c$.

$$\Phi(B) = \min_{\widetilde{\mathbf{B}}^c} \sum_{j=1}^{m} \sum_{i=1}^{k} |\beta_{ji} - \beta^c_{ji}|$$

Набор информационных матриц $\{I'_1, \ldots, I'_r\}$ назовём эквивалентным $\{I_1, \ldots, I_r\}$, если

$$I'_1 = \widetilde{K}(I_1), \ldots, I'_r = \widetilde{K}(I_r).$$

Задача оптимального коллективного синтеза сводится поиску эквивалентного $\{I_1, \ldots, I_r\}$ набора $\{I_1^m, \ldots, I_r^m\}$, для которого в результате применения сумматора получается матрица $\mathbf{B}_m$ с минимальным значением $\Phi$ среди всевозможных матриц , вычисляемых сумматором по наборам, эквивалентным $\{I_1, \ldots, I_r\}$.

При решении задач распознавания, классификации и анализа данных важное значение имеет наличие средств визуализации многомерных данных, позволяющих наглядно получать представление о конфигурации классов, кластеров и расположении отдельных объектов. Предполагаем опять, что у нас задана выборка $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$, состоящая из элементов пространства $\mathbb{R}^n$. Требуется найти отображение этого набора точек на плоскость $\mathbb{R}^2$ так, чтобы метрические соотношения между образами точек на плоскости максимально соответствовали бы метрическим соотношениям между ними в исходном признаковом пространстве.

- "Близкие"точки в исходном пространстве должны быть по-возможности "близкими"на плоскости.
- Соответственно "удалённые"точки в исходном пространстве должны быть по-возможности "удалёнными"на плоскости.

Пусть точки $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ являются образами точек $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ на плоскости $\mathbb{R}^2$. Пусть $\delta_{ij}$ - расстояние между векторами $\boldsymbol{x}_i$ и $\boldsymbol{x}_j$, $d_{ij}$ - расстояние между векторами $\boldsymbol{y}_i$ и $\boldsymbol{y}_j$. Ищется такое отображение, для которого сумма различий расстояний между точками будет минимальна

$$\mathbb{J}(\widetilde{\boldsymbol{y}}) = \sum_{i=1}^{m} \sum_{j=1}^{m} (\delta_{ij} - d_{ij})^2 \to \min,$$

где $\widetilde{\boldsymbol{y}} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ является вектором размерности $2m$, содержащим последовательно координаты двумерных образов многомерных объектов.
Минимизация функционала $\mathbb{J}(\widetilde{\boldsymbol{y}})$ проводится с помощью стандартной процедуры градиентного спуска.

Новое значение вектора координат двумерных образов $\widetilde{\boldsymbol{y}}^{l+1}$ на шаге $l$ вычисляется по значению вектора координат двумерных образов $\widetilde{\boldsymbol{y}}^{l1}$, вычисленному на предыдущем шаге, по формуле

$$\widetilde{\boldsymbol{y}}^{l+1} = \widetilde{\boldsymbol{y}}^{l} + \kappa \star \mathbf{grad}[\mathbb{J}(\widetilde{\boldsymbol{y}}^{l})]$$

где $\mathbf{grad}[\mathbb{J}(\widetilde{\boldsymbol{y}})]$ - градиент функционала $\mathbb{J}(\widetilde{\boldsymbol{y}})$ в точке $\widetilde{\boldsymbol{y}}$, $\kappa$ - шаг градиентного спуска. В качестве начальной конфигурации может использоваться проекция точек $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ на н плоскость, соответствующую некоторой паре признаков. Пример проекции на плоскость из пространства размерности 26, полученной описанным методом, приведён на рисунке 1.

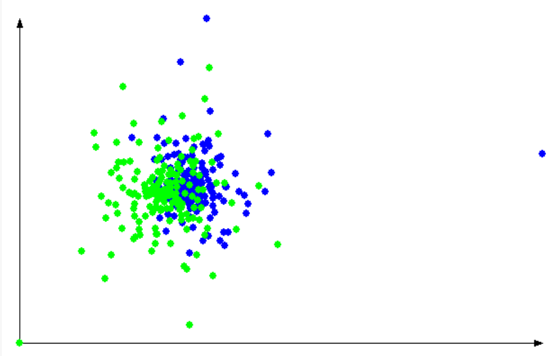

Рис.1. Точкам зелёного и синего цвета соответствуют описания двух классов объектов. Таким образом видно, что точки двух классов являются достаточно разделёнными.

# Методы преобразования признакового пространства. Метод главных компонент

Описанный метод многомерной визуализации фактически является методом нелинейного преобразования исходного признакового пространства. Вместе с тем существует эффективный метод линейной трансформации признакового пространства, позволяющий получить существенную информацию о структуре данных, а также получить новые признаки, удобные и эффективные при решении задач прогнозирования или распознавания. . Данный метод называется Методом главных компонент (Principal component analysis) , а также преобразованием Карунена-Лоэв (Karhunen–Loeve transform) Метод главных компонент основан на переходе от исходного множества вообще говоря коррелированных переменных $X_1, \ldots, X_n$ к новому набору переменных $Z_1, \ldots, Z_n$ с нулевыми коэффициентами ковариации между ними. То есть $cov(Z_i, Z_j) = 0$ при $i \neq j$, $i = 1, \ldots, n$, $j = 1, \ldots, n$. Переход к некоррелированным переменным может быть осуществлён с помощью линейного преобразования

Данное преобразование задаваётся матрицей вещественных коэффициентов $W = \|w_{ij}\|_{n\times n}$. Предположим, что у нас имеется исходная выборка $\widetilde{S}_{ini} = \{\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\}$ , которая может представлена в виде $\mathbf{X} = \begin{pmatrix} \mathrm{x}_{11} & \dots & \mathrm{x}_{1n} \\ \dots & \dots & \dots \\ \mathrm{x}_{j1} & \dots & \mathrm{x}_{jn} \\ \dots & \dots & \dots \\ \mathrm{x}_{m1} & \dots & \mathrm{x}_{mn} \end{pmatrix}$. . Далее будем предполагать,что признаки в матрице являются центрированными, то есть $\sum_{j=1}^{m} x_{ji} = 0$ при $i = 1,\dots,n$.Переход к центрированной выборке может быть всегда легко осуществлён с помощью простого линейного преобразования. Отметим, что $\mathbf{X}^t\mathbf{X} = (m-1)\boldsymbol{\Sigma}$, где $\boldsymbol{\Sigma} = \|\hat{\sigma_{ij}}\|_{n\times n}$ - выборочная ковариационная матрица, элементы которой вычисляются по формуле

$$\hat{\sigma_{ij}} = \frac{1}{m-1}\sum_{k=1}^{m} x_{ki}x_{kj}.$$

Подобная запись для коэффициентов ковариации возможна из за центрированности данных.

Предположим, что $\mathbf{Z} = \begin{pmatrix} \mathsf{z}_{11} & \dots & \mathsf{z}_{1n} \\ \dots & \dots & \dots \\ \mathsf{z}_{j1} & \dots & \mathsf{z}_{jn} \\ \dots & \dots & \dots \\ \mathsf{z}_{m1} & \dots & \mathsf{z}_{mn} \end{pmatrix}$. - матрица значения признаков $Z_1 \dots, Z_n$,

полученных с помощью линейного преобразования. Очевидно, что $\mathbf{Z} = \mathbf{XW}$ . Отметим, что матрица $\mathbf{Z}$ также является центрированной, поскольку линейное преобразование не приводит к утрате свойства центрированности. Вследствие требования отсутствия корреляции между переменными $Z_1, \dots, Z_n$ матрица $\mathbf{Z}^t\mathbf{Z}$ , являющаяся ковариационной матрицей для Z-переменных, является диагональной. При этом на диагонали в строке $i$ находится величина $(m-1)\delta_i$, где $\delta_i = \frac{1}{m-1}\sum_{j=1}^{m} z_{ji}^2$.

Однако

$$\mathbf{Z}^t\mathbf{Z} = \mathbf{W}^t\mathbf{X}^t\mathbf{X}\mathbf{W} = (m-1)\mathbf{W}^t\boldsymbol{\Sigma}\mathbf{W}.$$

Таким образом $\mathbf{W}^t\boldsymbol{\Sigma}\mathbf{W}$ является диагональной матрицей. Однако из теории матриц известно, что диагонализация симметрической вещественной матрицы $\boldsymbol{\Sigma}$ может быть осуществлена с помощью квадратной матрицы $\mathbf{E}$ , столбцами которой являются ортонормированные собственные вектора $\boldsymbol{\Sigma}$. Иными словами справедливо равенство $\mathbf{E}^t\boldsymbol{\Sigma}\mathbf{E} = \mathbf{V}$ , где $\mathbf{V}$ - диагональная матрица, на диагонали которой лежат собственные значения $\boldsymbol{\Sigma}$. Отсюда можно сделать вывод, что использование в качестве $\mathbf{W}$ матрицы $\mathbf{E}$ позволяет осуществлять переход к некоррелированным переменным $Z_1, \ldots, Z_n$. При этом дисперсиям переменных $Z_1, \ldots, Z_n$ будут соответствовать собственные значения матрицы $\boldsymbol{\Sigma}$ .

# Метод главных компонент

Столбцы матрицы $\mathbf{E}$ могут быть интерпретированы как новый ортонормированный базис в пространстве исходных переменных. При этом переменные $Z_1, \dots, Z_n$ являются проекциями на оси нового базиса. Отметим, что ось, соответствующая максимальному собственному значению, является одновременно тем направлением в исходном пространстве, для которого дисперсия проекций на него векторов обучающей выборки максимальна. Следует также отметить, что преобразование, задаваемое матрицей $\mathbf{E}$ является унитарным преобразованием, не изменяющим длины векторов. Вследствие этого полная выборочная дисперсия остаётся после преобразования $\mathbf{E}$ неизменной и равной сумме собственных значений ковариационной матрицы $\mathbf{\Sigma}$. То есть для полной дисперсии $D(\widetilde{S}_{ini})$ выборки $\widetilde{S}_{ini}$ справедливо разложение

$$D(\widetilde{S}_{ini}) = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{z}_j - \bar{\boldsymbol{z}})^2 \tag{1}$$

Однако из некоррелированности переменных $Z_1, \dots, Z_n$ следует, что

$$\frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{z}_j - \bar{\boldsymbol{z}})^2 = \sum_{i=1}^{n}\Lambda_i$$

где $\Lambda_1, \dots, \Lambda_n$ являются неотрицательными собственными значениями $\mathbf{\Sigma}$. В формуле (1) используются обозначения $\bar{\mathbf{x}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_j$, $\bar{\mathbf{z}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{z}_j$. Принимая во внимание условие центрированности переменных $X$, а значит и переменных $Z$, получаем

$$\frac{1}{m-1}\sum_{j=1}^{m}(\mathbf{z}_j)^2 = \sum_{i=1}^{n}\Lambda_i$$

Полученные в результате преобразования $\mathbf{E}$ переменные называют главными компонентами. Главные компоненты ранжируются в зависимости от величин соответствующих собственных значений.

Переменная, соответствующая максимальному собственному значению $\boldsymbol{\Lambda}_1$ и задаваемая соответствующим собственному вектором $\boldsymbol{e}_1$ называется первой главной компонентой. Она обладает максимальной дисперсией, равной $\boldsymbol{\Lambda}_1$. Переменная, соответствующая второму по величине собственному значению $\boldsymbol{\Lambda}_2$ и задаваемая соответствующим собственному вектором $\boldsymbol{e}_2$ называется второй главной компонентой и т.д.
Сумму $\sum_{i=1}^{k} \Lambda_i$ принято называть объяснённой дисперсией, а сумму $\sum_{i=k+1}^{n} \Lambda_i$ - остаточной дисперсией для $k$ первых главных компонент. Покажем справедливость равенства

$$\sum_{i=k+1}^{n} \Lambda_i = \frac{1}{m-1} \sum_{j=1}^{m} [\mathbf{x}_j - \sum_{i=1}^{k} \boldsymbol{e}_i(\boldsymbol{e}_i, \boldsymbol{x}_j)]^2.$$

Действительно, $\mathbf{x}_j = \sum_{i=1}^{n} \boldsymbol{e}_i(\boldsymbol{e}_i, \boldsymbol{x}_j)$. По определению переменных $Z$ справедливо равенство $\mathbf{x}_j = \sum_{i=1}^{n} \mathbf{e}_i z_{ji}$. Очевидно

$$\mathbf{x}_j - \sum_{i=1}^{k} \boldsymbol{e}_i(\boldsymbol{e}_i, \boldsymbol{x}_j) = \mathbf{x}_j - \sum_{i=1}^{k} \boldsymbol{e}_i z_{ji} = \sum_{i=k+1}^{n} \boldsymbol{e}_i z_{ji}$$

То есть

$$\frac{1}{m-1} \sum_{j=1}^{m} [\mathbf{x}_j - \sum_{i=1}^{k} \boldsymbol{e}_i(\boldsymbol{e}_i, \boldsymbol{x}_j)]^2 = \frac{1}{m-1} \sum_{j=1}^{m} ( \sum_{i=k+1}^{n} \boldsymbol{e}_i z_{ji})^2 =$$

$$= \frac{1}{m-1} \sum_{j=1}^{m} \sum_{i=k+1}^{n} {z_{ji}}^2.$$

Однако последняя сумма является суммой дисперсий переменных $Z_{k+1}, \ldots, Z_n$, равной остаточной дисперсии $\sum_{i=k+1}^{n} \Lambda_i$.

Иными словами остаточная дисперсия равна среднему квадрату расстоянию векторов до линейного замыкания k собственных векторов, соответствующих k первым главным компонентам. Таким образом низкая величина остаточной дисперсии соответствует хорошей аппроксимации данных этим линейным замыканием. На рисунке 2 представлен пример использования метода главных компонент для задачи с исходным признаковым пространством размерности 2. Собственный вектор, соответствующей первой главной (компоненте направлен вдоль прямой (c1,c2). Собственный вектор, соответствующей второй главной компоненте, перпендикулярен (c1,c2). Синими кружками отмечены точки, соответствующие данным. Из рисунка можно видеть значительную дисперсию проекций объектов выборки на направление первой главной компоненты. Следует отметить небольшую величину расстояний объектов обучающей выборки до прямой (c1,c2).

# Метод главных компонент

Отметим, что изначально метод главных компонент рассматривался К.Пирсоном как метод поиска линейных многообразий в признаковом пространстве, наилучшим образом аппроксимирующих данные. Метод главных компонент имеет широкий спектр применений, включая формирование нового признакового пространства, снижение размерности, визуализацию данных. Главные компоненты могут быть эффективно использованы при решении задач распознавания, регрессионного анализа. Метод очень широко используется в различных областях, включая биоинформатику, экономику, гуманитарные науки.

# Метод главных компонент

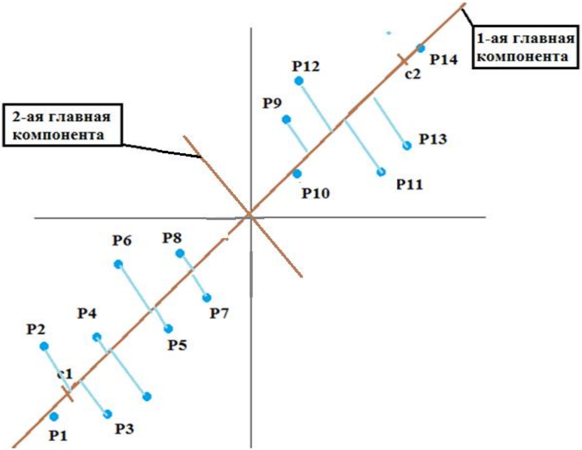


Рис. 2. пример использования метода главных компонент для задачи с исходным признаковым пространством размерности 2. Синими кружками отмечены точки, соответствующие данным.