ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Метод главных компонент

Высшая Школа Цифровой Культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Приветственное слово

Здравствуйте, уважаемые слушатели! Приветствуем вас в курсе «Продвинутое машинное обучение». В этом курсе мы разберем элементы факторного анализа на примере метода главных компонент, рассмотрим популярный и мощный метод классификации – метод опорных векторов (SVM), научимся вычислять информационную энтропию и на ее основе строить деревья принятия решений (ДПР), поговорим о том, как объединять модели в супермодель, рассмотрев ансамбли и закончим вишенкой на торте – обучением с подкреплением. Ну что, давайте начнем.

Работая с большим объемом данных неизвестной природы, можно почувствовать себя участником команды слепых, исследующих слона из известного примера[1]. И правда, как понять, есть ли в данных какие-либо закономерности и зависимости? Можно ли объединить объекты в какие-либо группы и так далее. На одномерных, двумерных или трехмерных данных ответ (или хотя бы намек на ответ) на эти вопросы часто позволяет дать экспертный опыт и человеческий глаз. В конце концов, если данные числовые, то их можно визуализировать и далее выдвигать какие-то предположения (тему нечисловых значений в этой лекции мы затрагивать не будем). А что, если размерность данных велика? Что тогда делать? С одной стороны, можно отбросить некоторые признаки и не включать их в рассмотрение. Такой подход иногда используется. Но тогда не понятно, не отказались ли мы от чего-то важного? С другой стороны, можно на основе имеющихся признаков синтезировать новые, количество которых будет не столь велико. Очевидно, что в таком случае, мы тоже потеряем некоторую часть информации о наших объектах, поэтому новые признаки должны быть максимально информативными. Но как этого добиться? Об этом и поговорим в этой лекции.

2 Наводящие размышления

Идея метода главных компонент (МГК) заключается в замене базиса с целью уменьшения размерности входных данных с минимальными потерями в информативности. Иными словами мы постараемся ввести новые предикторы для старых данных так, чтобы по новым предикторам информативность была максимальной. При этом мы не отбрасываем часть данных, а делаем некоторую композицию из признаков, которых в итоге становится меньше. Представляя данные в некоторой системе координат, мы можем ее поменять на ту, в которой данные больше всего отличаются по первой оси, далее, из оставшихся, больше всего отличаются по второй, и так далее. Идею легче

[1] https://www.datasciencecentral.com/profiles/blogs/the-story-of-big-data-data-science-amp-data-mining

всего понять из рисунка 1.

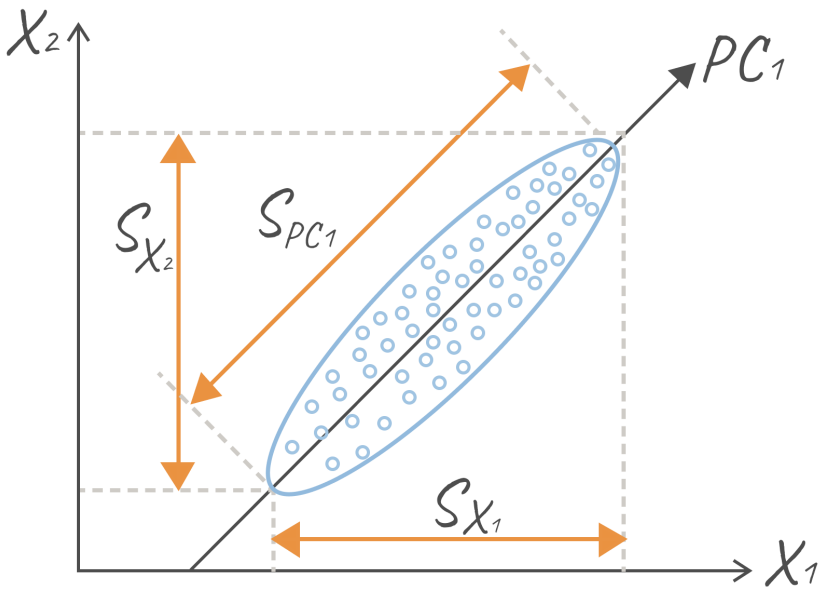

Рис. 1: Замена системы координат

Синие точки, образующие, условно, эллипс, – это данные (каждый объект обладает парой признаков). Вроде как логично, что вдоль прямой PC_1 изменение данных наиболее существенно, не так ли? Почему мы так считаем? А потому, что вдоль этой прямой, очевидно, разброс от среднего (условно, от центра эллипса) наибольший. В итоге, согласно нашей идее предполагается, что чем больше выборочная дисперсия вдоль оси, тем лучше (или больше, сильнее) меняются данные. Сама по себе идея построения новой системы координат не нова, она вторит идее и опыту в аналитической геометрии: выбери удобную систему координат, и задача решится в разы проще. Приведем еще один пример.

Пусть в компании есть пять менеджеров по продаже автомобилей. Руководителю отдела продаж необходимо распределить квартальную премию между этими сотрудниками в зависимости от количества проданных автомобилей. Под рассмотрение попадают два предиктора: количество автомобилей премиум сегмента и количество автомобилей эконом сегмента. Исходные данные представлены в таблице:

Сотрудник (x_i)	**Премиум** (X_1)**, шт**	**Эконом** (X_2)**, шт**
1	9	19
2	6	22
3	11	27
4	12	25
5	7	22

Для решения задачи о распределении премии удобно было бы, чтобы каждого сотрудника характеризовало одно единственное число, причем сотрудники должны быть максимально различимы, как этого добиться? Например, можно использовать некоторые коэффициенты φ_1 и φ_2, а итоговый «рейтинг» z_i сотрудника с номером i определять из соотношения

$$z_i = \varphi_1 x_{i1} + \varphi_2 x_{i2},\ i = \{1, 2, \ldots, 5\},$$

где x_{ij} – количество проданных автомобилей определенным сотрудником, индекс i отвечает за номер сотрудника (от 1 до 5), индекс j указывает на тип проданного автомобиля (1 – премиум, 2 – эконом). Например, $x_{52} = 22$ – это количество автомобилей класса эконом, проданных менеджером с номером 5. После описанного преобразования, каждого сотрудника с номером i будет характеризовать одно единственное число z_i, и по его значению можно решить задачу о распределении премии. Остается правильно подобрать эти коэффициенты φ_1 и φ_2, но чтобы это сделать, нам потребуется ввести терминологию и описать задачу математически.

3 Метод главных компонент

3.1 Идея метода на частном примере

3.1.1 Построение первой главной компоненты

Рассмотрим некоторый набор объектов $x_1, x_2, \ldots, x_n$, обладающих двумя признаками. Это значит, что каждый объект x_i можно отождествить с вектором, имеющим две координаты (x_{i1}, x_{i2}), то есть

$$x_i = (x_{i1}, x_{i2}),\quad i = \{1, 2, \ldots, n\}.$$

Имея такое представление, понятно, что множество этих объектов можно геометрически интерпретировать как множество точек на плоскости (рисунок 2).

Для начала выполним так называемое центрирование данных, то есть вычтем из каждой координаты каждого объекта среднее значение этой координаты по всем объектам. Это не влияет на идеологическую сторону описываемой ситуации (ведь мы просто-напросто сместили центр системы координат), но очень помогает в дальнейших преобразованиях. Результат представлен на рисунке 3.

После «схлопывания» исходного пространства признаков до одного измерения, каждому объекту x_i будет соответствовать единственная координата z_i:

$$x_i = (x_{i1}, x_{i2}) \longrightarrow z_i,\quad i \in \{1, 2, ..., n\}.$$

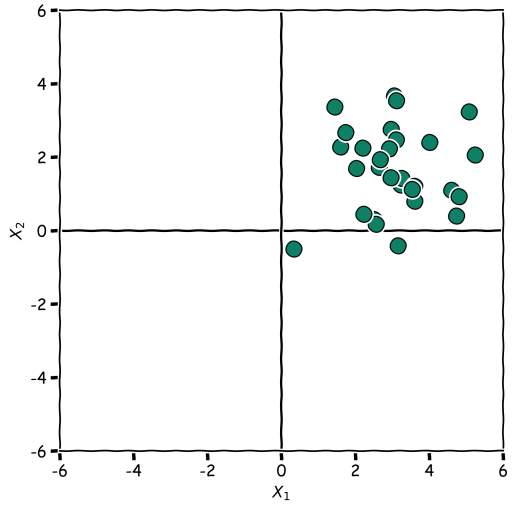

Рис. 2: Исходные данные до центрирования.

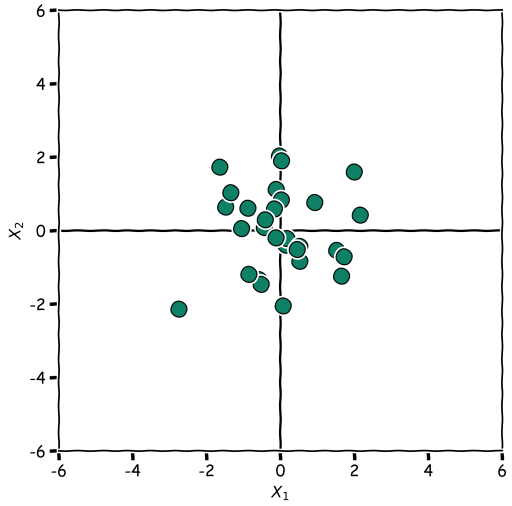

Рис. 3: Исходные данные после центрирования.

Тогда новые координаты всех объектов $x_1, x_2, ..., x_n$ можно записать в виде вектора-столбца

$$Z_1 = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}.$$

Если объект характеризуется единственным признаком, то его можно представить в виде точки на некоторой прямой. При этом, так как объекты центрированы, то разумно требовать, чтобы эта прямая проходила через начало координат, ведь тогда новые объекты останутся центрированными и относительно старой системы координат. Кроме того, мы хотим провести новую прямую так, чтобы новые координаты как можно сильнее различались между собой. В качестве меры различия при этом будем использовать выборочную дисперсию:

$$S^2(Z_1) = \frac{1}{n}\sum_{i=1}^{n}\left(z_i - \overline{Z_1}\right)^2,$$

где $\overline{Z_1} = \frac{1}{n}\sum_{i=1}^{n} z_i$ – выборочное среднее. В итоге оказывается разумным ввести следующее определение.

Определение 3.1.1 *Прямая, проходящая через начало координат, координаты проекций центрированных исходных объектов на которую обладают наибольшей выборочной дисперсией, называется первой главной компонентой и обозначается* PC_1.

Определение 3.1.2 *Направляющий вектор φ первой ГК, имеющий длину* 1*, называется вектором весов первой ГК.*

В случае двумерного пространства вектор весов φ будет иметь координаты φ_1 и φ_2.

Замечание 3.1.1 *Полезно заметить, что у каждой главной компоненты всегда существует ровно два вектора весов, один от другого отличающийся только направлением. Ясно, что на практике совершенно не важно, какой из них выбирается.*

Построим проекцию для одного объекта (рисунок 4), она обозначена синей точкой. На том же рисунке прямая – это первая главная компонента PC_1 с вектором весов $\varphi = (\varphi_1, \varphi_2)$, зеленая точка отвечает объекту x_i с координатами x_{i1} и x_{i2}, а значит, как уже было отмечено, может быть отождествлена с вектором $x_i = (x_{i1}, x_{i2})$.

Для нахождения новой координаты z_i на прямой PC_1 с базисным (в нашем случае совпадающим с вектором весов) вектором φ, можно воспользоваться скалярным произведением, которое определяется следующим образом:

$$(x_i, \varphi) = |x_i| \cdot |\varphi| \cos \alpha,$$

где α – угол между векторами x_i и φ. Учитывая, что $|\varphi| = 1$, получим

$$(x_i, \varphi) = |x_i| \cos \alpha = z_i,$$

То есть величина скалярного произведения (x_i, φ) дает величину z_i проекции x_i на направление φ – это и будет координатой на прямой PC_1 с базисным вектором φ. С другой стороны, доказывается, что в декартовой прямоугольной системе координат скалярное произведение может быть вычислено, как сумма произведений одноименных координат, а значит

$$z_i = (x_i, \varphi) = x_{i1}\varphi_1 + x_{i2}\varphi_2.$$

Аналогичную процедуру можно провести со всеми объектами рассматриваемого набора данных. Результаты представлены на рисунке 5. Проекции исходных объектов на прямую PC_1 отмечены синими точками.

Итак, получение проекций – дело не хитрое, а значит наша задача сводится к поиску таких весов φ_1 и φ_2, то есть к поиску такого способа проведения прямой, чтобы координаты проекций точек на эту прямую различались наиболее сильно, то есть обладали наибольшей выборочной дисперсией. Об этом мы поговорим чуть позже.

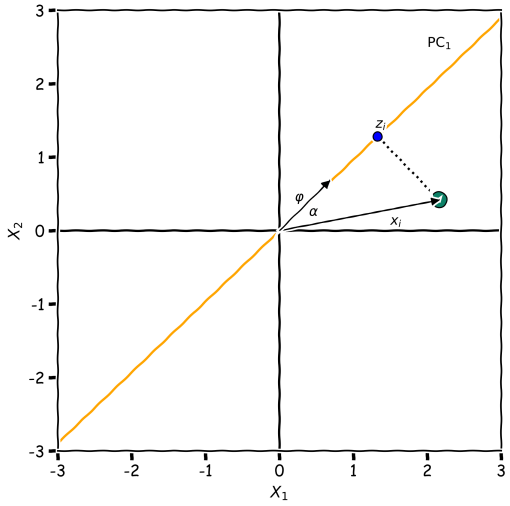

Рис. 4: Проекция объекта на первую ГК.

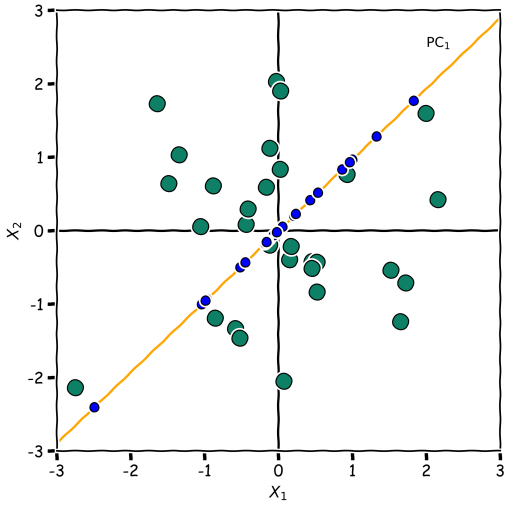

Рис. 5: Новые координаты объектов

3.1.2 Вторая и последующие главные компоненты

Давайте рассмотрим с точки зрения геометрии, что будет происходить в случае, если исходное пространство признаков имеет размерность $p \geq 2$ и нужно построить не одну, а k главных компонент, где $2 \leq k \leq p$. Отметим, что первая главная компонента во всех случаях строится аналогично рассмотренному примеру, то есть через начало координат проводится прямая так, чтобы координаты проекций точек на эту прямую обладали наибольшей выборочной дисперсией.

Вторая и последующие ГК строятся так, чтобы они также проходили через начало координат (ради сохранения центрированности объектов) ортогонально всем ранее построенным ГК. Требование ортогональности главных компонент обеспечивает отсутствие корреляции между новыми признаками объектов. При этом каждая последующая ГК строится так, чтобы выборочная дисперсия координат проекций исходных данных на нее была максимальной. На рисунках 6 и 7 представлены, соответственно, начальные данные, плоскость проектирования и проекция исходных объектов на плоскость.

3.2 Общее описание метода ГК

3.2.1 Постановка задачи и первая ГК

Теперь мы готовы рассмотреть строгое описание метода главных компонент в общем виде, а для этого сначала введем некоторые определения. При этом в данной лекции нам будет удобнее использовать матричные обозначения.

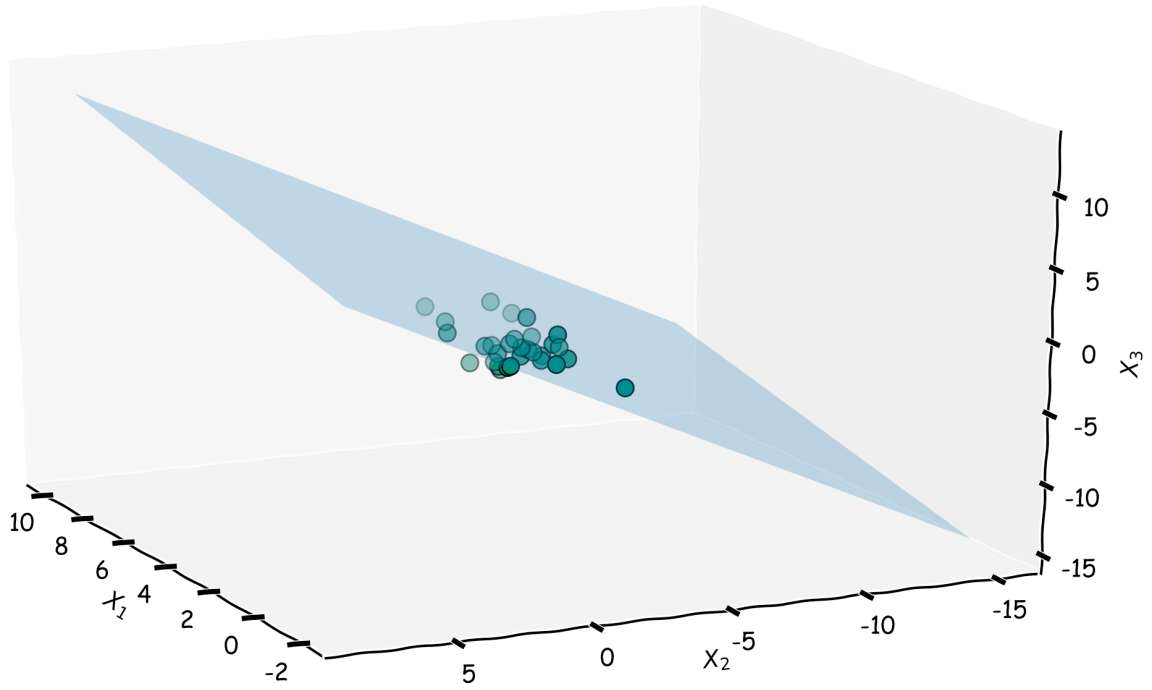

Рис. 6: Исходные данные и плоскость проектирования

Определение 3.2.1 *Пусть n объектов $x_1, x_2, \ldots, x_n$ имеет p признаков каждый, то есть*

$$\begin{aligned} x_1 &= \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \end{pmatrix}, \\ x_2 &= \begin{pmatrix} x_{21} & x_{22} & \ldots & x_{2p} \end{pmatrix}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ x_n &= \begin{pmatrix} x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}. \end{aligned}$$

Тогда матрицей исходных данных будем называть матрицу размера $[n \times p]$ вида:

$$F = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix},$$

где в i-ой строке матрицы F находятся соответствующие признаки i-го объекта.

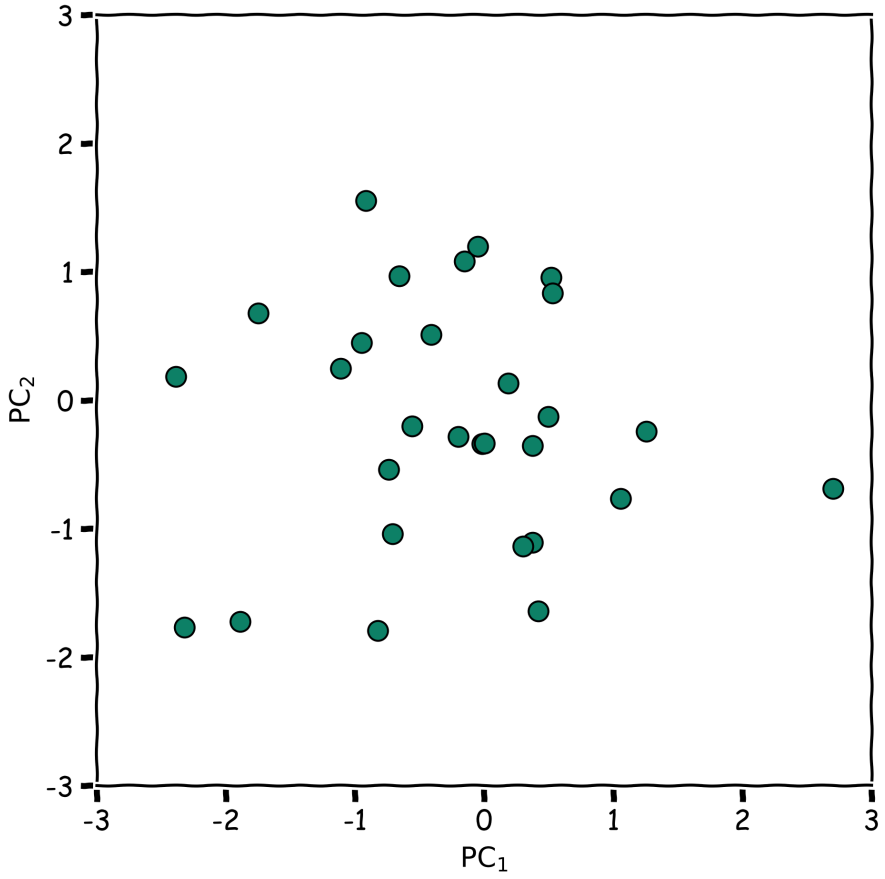

Рис. 7: Проекции исходных данных на плоскость

Здесь и далее будем считать, что матрица F получена в результате центрирования некоторой исходной матрицы объектов F'.

$$F' = \begin{pmatrix} x'_{11} & x'_{12} & \dots & x'_{1p} \\ x'_{21} & x'_{22} & \dots & x'_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x'_{n1} & x'_{n2} & \dots & x'_{np} \end{pmatrix}$$

Центрирование необходимо для того, чтобы каждый предиктор имел нулевое среднее значение и производится следующим образом: из каждого j-ого признака каждого объекта x_i вычитается среднее арифметическое того же j-ого признака, но взятое по всем объектам $x_1, x_2, ..., x_n$, то есть

$$x_{ij} = x'_{ij} - \overline{X'_j},\ i = \{1, 2, \dots, n\},\ j = \{1, 2, \dots, p\}.$$

где $\overline{X'_j}$ – среднее значение j-ого признака, то есть

$$\overline{X'_j} = \frac{x'_{1j} + x'_{2j} + \ldots + x'_{nj}}{n} = \frac{1}{n} \sum_{i=1}^{n} x'_{ij}.$$

Отметим отдельно, что последнее выражение является ни чем иным, как средним арифметическим элементов j-ого столбца матрицы F'. Покажем, что после описанного преобразования, выборочное среднее для всех признаков матрицы F действительно будет равно нулю. По определению, выборочное среднее по j-ому признаку – это среднее арифметическое $\overline{X_j}$ j-ого столбца матрицы F. Согласно тому, как выполнено центрирование, получим

$$\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{n}\sum_{i=1}^{n}(x'_{ij} - \overline{X'_j}) = \frac{1}{n}\sum_{i=1}^{n} x'_{ij} - \frac{1}{n}\sum_{i=1}^{n}\overline{X'_j} =$$

$$= \overline{X'_j} - \frac{1}{n}\cdot n\overline{X'_j} = 0.$$

Напомним также определение первой ГК и ее вектора весов.

Определение 3.2.2 *Прямая, проходящая через начало координат, координаты проекций центрированных исходных объектов на которую обладают наибольшей выборочной дисперсией, называется первой главной компонентой и обозначается* PC_1.

Определение 3.2.3 *Направляющий вектор* $\varphi_1 = (\varphi_{11}, \varphi_{21}, \ldots, \varphi_{p1})$ *первой главной компоненты, имеющий длину* 1, *называется вектором весов первой главной компоненты.*

Так как длина вектора φ_1 равна единице, то есть

$$|\varphi_1| = \sqrt{\varphi_{11}^2 + \varphi_{21}^2 + \ldots + \varphi_{p1}^2} = 1,$$

то и

$$\varphi_{11}^2 + \varphi_{21}^2 + \ldots + \varphi_{p1}^2 = 1.$$

В матричных обозначениях будем записывать координаты вектора φ_1 в виде одноименного вектора-столбца

$$\varphi_1 = \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \\ \vdots \\ \varphi_{p1} \end{pmatrix}.$$

Замечание 3.2.1 *Полезно задуматься, из каких соображений вводится ограничение на длину вектора весов. Если длина вектора весов равна единице, то, как было показано ранее, координата проекции объекта на главную компоненту с базисным вектором – вектором весов, может быть получена, как результат скалярного произведения между соответствующим этому объекту вектором и вектором весов. В итоге, ограничение на длину вектора весов позволяет упростить вычисления.*

Определение 3.2.4 *Вектором счётов первой ГК называется вектор-столбец Z_1 вида*

$$Z_1 = \begin{pmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{pmatrix},$$

состоящий из координат проекций центрированных исходных данных на вектор весов первой ГК.

Иными словами, вектор счётов первой ГК – это новые координаты объектов относительной первой главной компоненты.

Итак, как мы уже отмечали, если вектор весов φ_1 первой главной компоненты найден, то мы можем вычислить и вектор счётов первой главной компоненты следующим образом.

Теорема 3.2.1 *Пусть*

$$F = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

– матрица, состоящая из центрированных исходных данных, и φ_1 – вектор весов первой главной компоненты. Тогда вектор счётов первой ГК может быть получен из соотношения

$$Z_1 = F\varphi_1.$$

Доказательство. Мы уже отмечали, что в случае, когда вектор весов первой главной компоненты $\varphi_1 = (\varphi_{11}, \varphi_{21}, \dots, \varphi_{p1})$ имеет единичную длину, то координата проекции z_{i1} произвольного объекта $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ может быть вычислена, как

$$z_{i1} = x_{i1}\varphi_{11} + x_{i2}\varphi_{21} + \dots + x_{ip}\varphi_{p1},$$

что, если внимательно посмотреть на матричное представление заявленного соотношения

$$Z_1 = F\varphi_1 = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \\ \vdots \\ \varphi_{p1} \end{pmatrix},$$

и есть не что иное, как i-ая координата вектора Z_1, которая получается согласно правилам матричного умножения: произведение i-ой строки на столбец. Это может быть записано следующим образом:

$$Z_1 = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \\ \vdots \\ \varphi_{p1} \end{pmatrix} =$$

$$= \begin{pmatrix} x_{11}\varphi_{11} + x_{12}\varphi_{21} + \dots + x_{1p}\varphi_{p1} \\ x_{21}\varphi_{11} + x_{22}\varphi_{21} + \dots + x_{2p}\varphi_{p1} \\ \vdots \\ x_{n1}\varphi_{11} + x_{n2}\varphi_{21} + \dots + x_{np}\varphi_{p1} \end{pmatrix} = \begin{pmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{pmatrix}.$$

□

3.2.2 Нахождение первой ГК

Нами так и не решен вопрос: как же все-таки искать первую главную компоненту? Этот вопрос сводится к вопросу: а как искать вектор весов? Давайте разбираться!

Как мы только что установили, вектор счётов Z_1 первой главной компоненты PC_1 – это произведение матрицы данных на вектор весов φ_1:

$$Z_1 = F\varphi_1,$$

то есть

$$\begin{pmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \\ \vdots \\ \varphi_{p1} \end{pmatrix},$$

при условии, что выборочная дисперсия полученного набора максимальна. Напомним, что выборочная дисперсия выборки Z_1 может быть найдена по следующей формуле:

$$S^2(Z_1) = \overline{Z_1^2} - \overline{Z_1}^2,$$

где $\overline{Z_1^2}$ – выборочное среднее квадратов значений Z_1, а $\overline{Z_1}^2$ – квадрат выборочного среднего. В координатах получим

$$S^2(Z_1) = \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}\right)^2.$$

Рассмотрим правую часть этой разности, а точнее выражение в скобках. Как мы уже отмечали, согласно правилу умножения матриц,

$$z_{i1} = \varphi_{11}x_{i1} + \varphi_{21}x_{i2} + \cdots + \varphi_{p1}x_{ip},\ i = \{1, 2, \ldots, n\}.$$

Тогда

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1} = \frac{1}{n}\sum_{i=1}^{n}(\varphi_{11}x_{i1} + \varphi_{21}x_{i2} + \cdots + \varphi_{p1}x_{ip}) =$$

$$= \varphi_{11}\frac{1}{n}\sum_{i=1}^{n} x_{i1} + \varphi_{21}\frac{1}{n}\sum_{i=1}^{n} x_{i2} + \cdots + \varphi_{p1}\frac{1}{n}\sum_{i=1}^{n} x_{ip}.$$

С учетом того, что признаки x_{ij} центрированы, а значит их выборочные средние $\overline{X_j}$ равны нулю, то есть

$$\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0,\ j = \{1, 2, \ldots, p\},$$

то мы получаем, что каждое слагаемое в последней сумме равно нулю, значит вся сумма равна нулю, а тогда и квадрат этой суммы тоже равен нулю:

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1} = 0 \Longrightarrow \left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}\right)^2 = 0.$$

Возвращаясь к выборочной дисперсии $S^2(Z_1)$, получим, что максимизировать нужно следующее выражение:

$$S^2(Z_1) = \frac{1}{n}\sum_{i=1}^{n} z_{i1}^2 \longrightarrow \max_{\varphi_1},$$

при условии, что $|\varphi_1| = 1$. Так как в выражении мы вольны менять только координаты φ_1, то на самом деле задача сводится к нахождению вектора весов φ_1 при котором $S^2(Z_1)$ будет максимальной. Коротко математическим языком задачу можно записать следующим образом:

$$\arg\max_{\varphi_1}\left(\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2\right) = \arg\max_{\varphi_1}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\varphi_{j1}x_{ij}\right)^2\right),$$

при условии, что $|\varphi_1| = 1$. Так как задача оптимизации не зависит от n, то оказывается, что максимизировать нужно квадрат длины вектора Z_1. Иными словами, мы ищем такой вектор весов, что квадрат длины $|Z_1|^2$ вектора

счётов Z_1 максимален. На математическом языке задача формулируется так:

$$\arg\max_{\varphi_1} \left(\sum_{i=1}^{n} z_{i1}^2 \right) = \arg\max_{\varphi_1} \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \varphi_{j1} x_{ij} \right)^2 \right),$$

при условии, что $|\varphi_1| = 1$.

3.2.3 Нахождение произвольной ГК

Итак, мы поняли, что такое первая главная компонента, ее вектор весов и вектор счётов. Но что, если мы хотим сопоставить каждому объекту больше, чем одну координату? Введем следующее определение.

Определение 3.2.5 *Пусть построена первая главная компонента* PC_1, $p \geq 2$. *Прямая, проходящая через начало координат ортогонально первой главной компоненте* PC_1, *координаты проекций центрированных исходных объектов на которую обладают наибольшей выборочной дисперсией, называется второй главной компонентой и обозначается* PC_2.

И вообще, пусть построено $k-1$ *главных компонент* $\mathsf{PC}_1, \mathsf{PC}_2, ..., \mathsf{PC}_{k-1}$. *Прямая, проходящая через начало координат ортогонально каждой главной компоненте* PC_i, $i \in \{1, 2, ..., (k-1)\}$, $k-1 < p$, *координаты проекций центрированных исходных объектов на которую обладают наибольшей выборочной дисперсией, называется* k*-ой главной компонентой и обозначается* PC_k.

Итак, каждая последующая главная компонента – это прямая, проходящая через начало координат ортогонально всем ранее построенным главным компонентам так, чтобы координаты проекций объектов на нее имели наибольшую выборочную дисперсию. Требование ортогональности объясняется, как было сказано ранее, отсутствием корреляции новых координат. Кроме того, количество главных компонент не может быть больше, чем размерность исходного пространства, то есть не может быть больше p. Подумайте, почему так?

Аналогично тому, как было сделано ранее, введем определения весов.

Определение 3.2.6 *Направляющий вектор* $\varphi_k = (\varphi_{1k}, \varphi_{2k}, ..., \varphi_{pk})$ k*-ой главной компоненты, имеющий длину* 1, *называется вектором весов* k*-ой главной компоненты.*

В матричных обозначениях будем записывать координаты вектора φ_k в виде

одноименного вектора-столбца:

$$\varphi_k = \begin{pmatrix} \varphi_{1k} \\ \varphi_{2k} \\ \vdots \\ \varphi_{pk} \end{pmatrix}.$$

Замечание 3.2.2 *Для удобства дальнейшего описания алгоритма поиска главных компонент отметим, что ортогональность главных компонент эквивалентна ортогональности их направляющих векторов. В координатах ортогональность векторов φ_i и φ_j при $i \neq j$ может быть записана следующим образом:*

$$\varphi_{1i}\varphi_{1j} + \varphi_{2i}\varphi_{2j} + \ldots + \varphi_{pi}\varphi_{pj} = 0.$$

Последнее равенство есть не что иное, как равенство нулю скалярного произведения векторов $\varphi_i = (\varphi_{1i}, \varphi_{2i}, ..., \varphi_{pi})$ и $\varphi_j = (\varphi_{1j}, \varphi_{2j}, ..., \varphi_{pj})$.

Наша цель, конечно, не столь главные компоненты, сколько счёты, поэтому введем финальное определение.

Определение 3.2.7 *Вектором счётов k-ой главной компоненты называется вектор Z_k вида*

$$Z_k = \begin{pmatrix} z_{1k} \\ z_{2k} \\ \vdots \\ z_{nk} \end{pmatrix},$$

состоящий из координат проекций центрированных исходных данных на вектор весов k-ой главной компоненты.

Как и ранее, зная вектор весов, вектор счётов может быть получен следующим образом.

Теорема 3.2.2 *Пусть*

$$F = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

– матрица, состоящая из центрированных исходных данных, и φ_k – вектор весов k-ой главной компоненты. Тогда вектор счётов k-ой главной компоненты может быть получен из соотношения

$$Z_k = F\varphi_k.$$

Не вдаваясь в детали, которые подробно обсуждались ранее, при поиске k-ой главной компоненты решается задача максимизации квадрата длины $|Z_k|^2$ вектора счётов Z_k при условии, что вектор весов φ_k ортогонален всем ранее построенным векторам весов $\varphi_1, \varphi_2, ..., \varphi_{k-1}$. Более коротко эта задача записывается следующим образом:

$$\arg\max_{\varphi_k} \left(\sum_{i=1}^{n} z_{ik}^2 \right) = \arg\max_{\varphi_k} \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \varphi_{jk} x_{ij} \right)^2 \right),$$

$$\varphi_k \perp \varphi_i,\ i \in \{1, 2, ..., (k-1)\},\ |\varphi_k| = 1$$

Резюмируя, поиск k-ой главной компоненты или, что то же самое, вектора весов φ_k, и, как следствие, вектора счётов Z_k, осуществляется по следующему алгоритму.

1. Постулируется, что вектор φ_k имеет единичную длину, то есть

$$\sum_{i=1}^{p} \varphi_{ik}^2 = \varphi_{1k}^2 + \varphi_{2k}^2 + ... + \varphi_{pk}^2 = 1.$$

2. Постулируется, что вектор φ_k ортогонален каждому из векторов весов $\varphi_1, \varphi_2, ..., \varphi_{k-1}$. В координатах это эквивалентно $(k-1)$-ому равенству вида

$$\varphi_{1k}\varphi_{1i} + \varphi_{2k}\varphi_{2i} + ... + \varphi_{pk}\varphi_{pi} = 0,\ i \in \{1, 2, ..., (k-1)\}$$

3. При объявленных условиях решается задача

$$\arg\max_{\varphi_k} \left(|Z_k|^2 \right) = \arg\max_{\varphi_k} \left(\sum_{i=1}^{n} z_{ik}^2 \right) = \arg\max_{\varphi_k} \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \varphi_{jk} x_{ij} \right)^2 \right).$$

4. После чего для вычисления вектора счётов k-ой главной компоненты вычисляется

$$Z_k = F\varphi_k.$$

3.3 Пример нахождения первой главной компоненты

Вернемся к нашему примеру с продавцами автомобилей. Напомним исходные данные, которые приведены в таблице.

Сотрудник (x_i)	**Премиум** (X'_1), **шт**	**Эконом** (X'_2), **шт**
1	9	19
2	6	22
3	11	27
4	12	25
5	7	22

Найдем средние значения каждого признака:

$$\overline{X'_1} = \frac{9+6+11+12+7}{5} = 9,$$

$$\overline{X'_2} = \frac{19+22+27+25+22}{5} = 23$$

и выполним центрирование для объекта x_1

$$x_{11} = x'_{11} - \overline{X'_1} = 9 - 9 = 0, \quad x_{12} = x'_{12} - \overline{X'_2} = 19 - 23 = -4,$$

а также для объекта x_2

$$x_{21} = x'_{21} - \overline{X'_1} = 6 - 9 = -3, \quad x_{22} = x'_{22} - \overline{X'_2} = 22 - 23 = -1,$$

и так далее. В результате получим таблицу координат объектов после центрирования.

Сотрудник (x_i)	**Признак** X_1	**Признак** X_2
1	0	-4
2	-3	-1
3	2	4
4	3	2
5	-2	-1

По написанной таблице составим матрицу центрированных исходных данных:

$$F_{5\times 2} = \begin{pmatrix} 0 & -4 \\ -3 & -1 \\ 2 & 4 \\ 3 & 2 \\ -2 & -1 \end{pmatrix}.$$

Напомним, что Z_1 ищется в виде:

$$Z_1 = F\varphi_1,$$

где $\varphi_1 = \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \end{pmatrix}$ – вектор весов, а значит

$$\varphi_{11}^2 + \varphi_{21}^2 = 1.$$

Согласно описанному алгоритму, для вычисления счётов первой главной компоненты достаточно вычислить

$$Z_1 = \begin{pmatrix} 0 & -4 \\ -3 & -1 \\ 2 & 4 \\ 3 & 2 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \end{pmatrix} = \begin{pmatrix} -4\varphi_{21} \\ -3\varphi_{11} - \varphi_{21} \\ 2\varphi_{11} + 4\varphi_{21} \\ 3\varphi_{11} + 2\varphi_{21} \\ -2\varphi_{11} - \varphi_{21} \end{pmatrix}$$

Как мы уже отмечали, для максимизации выборочной дисперсии $S^2(Z_1)$ достаточно максимизировать квадрат длины Z_1, учитывая, что $|\varphi_1| = 1$, то есть максимизировать выражение

$$|Z_1|^2 = (-4\varphi_{21})^2 + (-3\varphi_{11} - \varphi_{21})^2 + (2\varphi_{11} + 4\varphi_{21})^2 +$$

$$+ (3\varphi_{11} + 2\varphi_{21})^2 + (-2\varphi_{11} - \varphi_{21})^2 .$$

Раскроем скобки и приведем подобные слагаемые. В результате получим выражение для квадрата длины Z_1 следующего вида:

$$|Z_1|^2 = 26\varphi_{11}^2 + 38\varphi_{11}\varphi_{21} + 38\varphi_{21}^2.$$

Перед нами функция двух переменных, максимальное значение которой при условии $|\varphi_1| = 1$ можно найти, используя метод Лагранжа, однако, можно использовать следующий прием. Рассмотрим 2 случая:

1. Так как

$$\varphi_{11}^2 + \varphi_{21}^2 = 1,$$

то выражение для $|Z_1|^2$ можно переписать в виде:

$$|Z_1|^2 = 26 + \frac{38\varphi_{11}\varphi_{21}}{\varphi_{11}^2 + \varphi_{21}^2} + \frac{12\varphi_{21}^2}{\varphi_{11}^2 + \varphi_{21}^2}.$$

Пусть $\varphi_{11} \neq 0$. Разделим в каждой дроби числитель и знаменатель на φ_{11}^2 и введем замену $t = \frac{\varphi_{21}}{\varphi_{11}}$. При этом t может принимать любые значения из множества действительных чисел. Рассмотрим функцию одной переменной $G(t)$

$$G(t) = 26 + \frac{12t^2}{1 + t^2} + \frac{38t}{1 + t^2},$$

максимальное значение которой мы хотим найти. Напомним, что для нахождения точек, подозрительных на экстремум, можно найти первую производную и найти нули как числителя, так и знаменателя. Тогда

$$G'(t) = \frac{-38t^2 + 24t + 38}{(1+t^2)^2}.$$

Так как знаменатель в ноль не обращается, то приравняем к нулю лишь числитель, откуда придем к квадратному уравнению

$$-38t^2 + 24t + 38 = 0,$$

с корнями

$$t_{1,2} = \frac{6 \pm \sqrt{397}}{19}.$$

Функция имеет локальный максимум в точке, где производная меняет знак с плюса на минус. В нашем случае $t = \frac{6+\sqrt{397}}{19}$ – точка локального максимума, причем

$$G\left(\frac{6+\sqrt{397}}{19}\right) = 32 + \sqrt{397} \approx 51.925.$$

С другой стороны, функция $G(t)$ убывает на интервале $\left(-\infty, \frac{6-\sqrt{397}}{19}\right)$, а значит стремится «принять» свое наибольшее значение при $t \to -\infty$. Для нахождения этого значения вычислим предел

$$\lim_{t\to-\infty}\left(26 + \frac{12t^2}{1+t^2} + \frac{38t}{1+t^2}\right) = 38.$$

2. Теперь рассмотрим случай $\varphi_{11} = 0$. Тогда, учитывая, что

$$|Z_1|^2 = 26\varphi_{11}^2 + 38\varphi_{11}\varphi_{21} + 38\varphi_{21}^2,$$

и

$$\varphi_{11}^2 + \varphi_{21}^2 = 1,$$

получим, что $\varphi_{21} = \pm 1$ и

$$|Z_1|^2 = 38.$$

Сравнивая все три полученных значения, очевидно, что наибольшее значение рассматриваемой функции достигается при

$$\frac{\varphi_{21}}{\varphi_{11}} = \frac{6+\sqrt{397}}{19}.$$

Тогда, из системы уравнений

$$\begin{cases} \varphi_{11}^2 + \varphi_{21}^2 = 1 \\ \frac{\varphi_{21}}{\varphi_{11}} = \frac{6+\sqrt{397}}{19} \end{cases},$$

получим 2 пары решений: $\varphi_{11} \approx 0.591$, $\varphi_{21} \approx 0.807$ и $\varphi_{11} \approx -0.591$, $\varphi_{21} \approx -0.807$. Какую пару выбрать, значения не имеет, так как отличие только в направлении. В таком случае новые координаты (или счёты) объектов, относительно первой ГК, можно записать в виде:

$$Z_1 = \begin{pmatrix} 0 & -4 \\ -3 & -1 \\ 2 & 4 \\ 3 & 2 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} 0.591 \\ 0.807 \end{pmatrix} = \begin{pmatrix} -3.226 \\ -2.580 \\ 4.409 \\ 3.387 \\ -1.989 \end{pmatrix}.$$

Так как направление первой ГК задается вектором весов φ_1, то можно построить прямую, имеющую уравнение:

$$X_2 = \frac{\varphi_{21}}{\varphi_{11}} X_1 = \frac{0.807}{0.591} X_1 = 1.365 X_1.$$

Геометрическая интерпретация представлена на рисунке **??**. Зеленые точки – центрированные исходные объекты. Оранжевая линия – первая ГК, то есть прямая с вектором весов φ_1. Новые координаты (или счёты) – это координаты проекций исходных объектов на полученную прямую (синие точки). При этом, в соответствии с описанием метода, прямая выбрана таким образом, чтобы выборочная дисперсия счётов была наибольшей.

Теперь каждого менеджера характеризует одно число, а значит если упорядочить результаты по убыванию, наибольшую премию должен получить сотрудник с номером $i = 3$, затем сотрудник с номером $i = 4$ и так далее. Заметим, что отрицательные результаты не значат, что сотрудников нужно оштрафовать, просто чем меньше значение, тем меньшей премии достоен сотрудник.

4 Еще один способ построения главных компонент

Как мы уже установили, поиск главных компонент – весьма полезная, но трудоемкая задача, особенно учитывая то, что при поиске каждой последующей компоненты необходимо решать оптимизационную задачу, со все большим числом ограничений (ортогональность вектора весов последующей ГК всем векторам весов предыдущих ГК). С другой стороны, возникает ряд справедливых вопросов. А не мало ли построено ГК? Не потеряли ли мы слишком

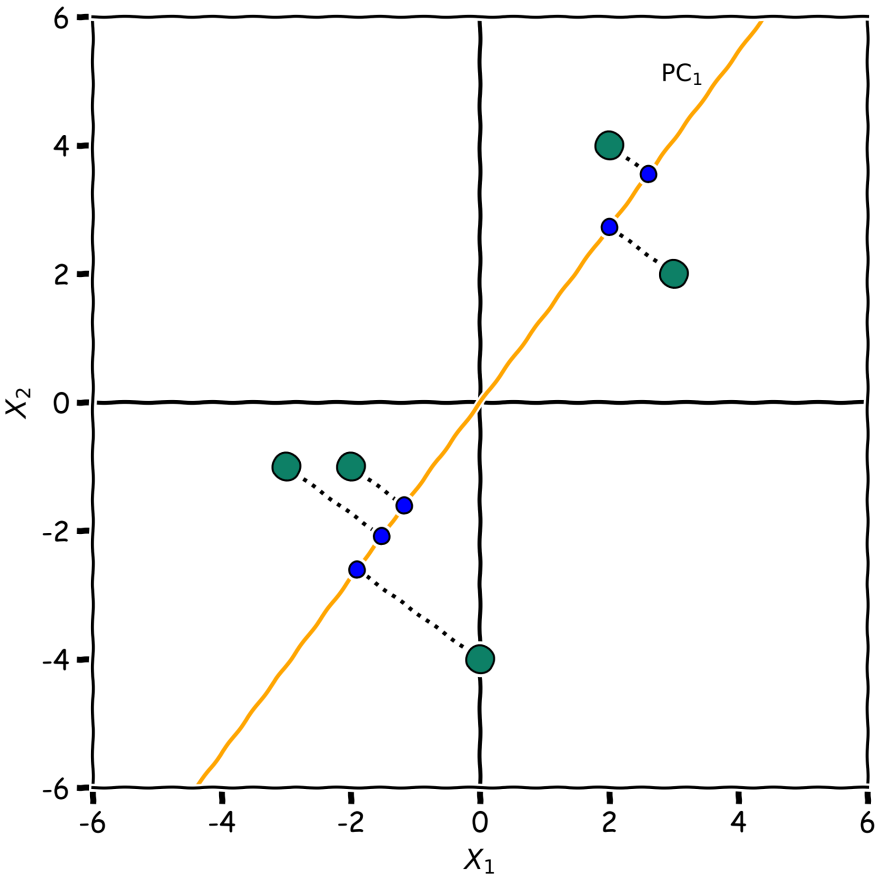

Рис. 8: Построение первой ГК

много полезной информации о рассматриваемых объектах в ходе снижения размерности пространства признаков? Можно ли как-то посчитать, какую часть информации мы потеряли? Можно ли как-то восстановить (пусть и с потерями) исходные данные? Ну и существует ли более общий метод, позволяющий находить векторы весов главных компонент не итеративно, по очереди, а сразу все? На все эти вопросы можно ответить утвердительно, но для этого придется провести подготовительную работу.

4.1 Некоторые сведения из линейной алгебры

Как вы, наверное, уже поняли, и как мы отмечали в самом начале, при использовании метода главных компонент мы, по сути дела, меняем систему координат, или, что то же самое, базис. А как в математике в принципе происходит замена базиса? Давайте вспомним.

Пусть заданы два базиса. Первый состоит из элементов $e_1, e_2, \ldots, e_p$ (условно назовем его старым), а второй – из элементов $e'_1, e'_2, \ldots, e'_p$ (условно назовем его новым). Так как перед нами два базиса в одном и том же пространстве, то векторы нового базиса могут быть выражены через векторы старого следующим образом:

$$e'_1 = \varphi_{11}e_1 + \varphi_{21}e_2 + \cdots + \varphi_{p1}e_p,$$
$$e'_2 = \varphi_{12}e_1 + \varphi_{22}e_2 + \cdots + \varphi_{p2}e_p,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$e'_p = \varphi_{1p}e_1 + \varphi_{2p}e_2 + \cdots + \varphi_{pp}e_p,$$

где φ_{ij} – числа, $i, j \in \{1, 2, ..., p\}$. Оказывается, для различных преобразований координат удобно составить так называемую матрицу перехода от старого базиса к новому.

Определение 4.1.1 *Матрица*

$$\Phi = \begin{pmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1p} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{p1} & \varphi_{p2} & \cdots & \varphi_{pp} \end{pmatrix}$$

называется матрицей перехода от базиса $e_1, e_2, ..., e_p$ *к базису* $e'_1, e'_2, ..., e'_p$.

Обратите внимание, что коэффициенты разложения вектора e'_1 по старому базису стоят в первом столбце матрицы перехода Φ, коэффициенты разложения вектора e'_2 по старому базису стоят во втором столбце матрицы перехода Φ, ну и так далее.

С помощью таким образом составленной матрицы Φ очень удобно переводить координаты произвольного вектора из нового базиса в старый.

Лемма 4.1.1 *Пусть в базисах* $e_1, e_2, ..., e_p$ *и* $e'_1, e'_2, ..., e'_p$ *вектор* X *задается, как*

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \text{и} \quad X = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_p \end{pmatrix},$$

соответственно, а Φ *– матрица перехода от старого базиса к новому. Тогда справедливо следующее соотношение:*

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \Phi \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_p \end{pmatrix}.$$

Итак, координаты вектора X в старом базисе – это произведение матрицы перехода на координаты вектора X в новом базисе.

Доказательство. Запишем разложение вектора X в новом и старом базисах. Получим, с одной стороны,

$$X = x_1e_1 + x_2e_2 + \ldots + x_pe_p,$$

а с другой –

$$X = x'_1e'_1 + x'_2e'_2 + \ldots + x'_pe'_p.$$

Так как новый и старый базисы связаны соотношениями,

$$e'_1 = \varphi_{11}e_1 + \varphi_{21}e_2 + \cdots + \varphi_{p1}e_p,$$

$$e'_2 = \varphi_{12}e_1 + \varphi_{22}e_2 + \cdots + \varphi_{p2}e_p,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$e'_p = \varphi_{1p}e_1 + \varphi_{2p}e_2 + \cdots + \varphi_{pp}e_p,$$

то, выполнив подстановку и перегруппировав слагаемые в правой части равенства, собрав подобные слагаемые перед базисными элементами e_i, получим

$$X = x'_1e'_1 + x'_2e'_2 + \ldots + x'_pe'_p = (x'_1\varphi_{11} + x'_2\varphi_{12} + \ldots + x'_p\varphi_{1p})e_1 +$$

$$+(x'_1\varphi_{21} + x'_2\varphi_{22} + \ldots + x'_p\varphi_{2p})e_2 + \ldots + (x'_1\varphi_{p1} + x'_2\varphi_{p2} + \ldots + x'_p\varphi_{pp})e_p.$$

С другой стороны, как уже отмечалось,

$$X = x_1e_1 + x_2e_2 + \ldots + x_pe_p.$$

Воспользовавшись единственностью разложения вектора по базису, имеем

$$x_1 = x'_1\varphi_{11} + x'_2\varphi_{12} + \ldots + x'_p\varphi_{1p},$$

$$x_2 = x'_1\varphi_{21} + x'_2\varphi_{22} + \ldots + x'_p\varphi_{2p},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_p = x'_1\varphi_{p1} + x'_2\varphi_{p2} + \ldots + x'_p\varphi_{pp}.$$

В матричном виде это можно представить как

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} \varphi_{11} & \varphi_{12} & \ldots & \varphi_{1p} \\ \varphi_{21} & \varphi_{22} & \ldots & \varphi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{p1} & \varphi_{p2} & \ldots & \varphi_{pp} \end{pmatrix} \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_p \end{pmatrix}.$$

□

В МГК мы меняем не произвольный базис на произвольный, а ортонормированный на ортонормированный. Давайте вспомним, что это значит.

Определение 4.1.2 *Базис* $e_1, e_2, ..., e_p$ *называется ортонормированным, если*

$$|e_1| = |e_2| = \ldots = |e_p| = 1$$

и элементы e_i *и* e_j *ортогональны при* $i \neq j$, *то есть* $(e_i, e_j) = 0$, $i \neq j$.

Говоря проще, базис называется ортонормированным, если длины всех его элементов равны единице, а элементы попарно ортогональны. Заметьте, это ровно-таки те требования к векторам весов, которые мы ищем при использовании МГК.

В случае замены ортонормированного базиса на ортонормированный оказывается, что матрица перехода Φ обладает полезными для дальнейшего свойствами.

Теорема 4.1.1 *Пусть* $e_1, e_2, ..., e_p$ *и* $e'_1, e'_2, ..., e'_p$ *– два ортонормированных базиса и* Φ *– матрица перехода. Тогда*

I. *Столбцы матрицы* Φ *имеют единичную длину и попарно ортогональны.*

II. *Строки матрицы* Φ *имеют единичную длину и попарно ортогональны.*

III. $\Phi^{-1} = \Phi^T$, *то есть обратная и транспонированная матрицы совпадают.*

Доказательство. Докажем сначала пункты I и III, а затем пункт II.
I. Воспользуемся формулами перехода из одного базиса в другой, а именно

$$e'_1 = \varphi_{11}e_1 + \varphi_{21}e_2 + \cdots + \varphi_{p1}e_p,$$

$$e'_2 = \varphi_{12}e_1 + \varphi_{22}e_2 + \cdots + \varphi_{p2}e_p,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$e'_p = \varphi_{1p}e_1 + \varphi_{2p}e_2 + \cdots + \varphi_{pp}e_p.$$

Отметим, что так как базисы $e_1, e_2, ..., e_p$ и $e'_1, e'_2, ..., e'_p$ ортонормированы, то

$$(e'_k, e'_l) = (e_k, e_l) = \begin{cases} 1, & k = l \\ 0, & k \neq l \end{cases}, \quad k,\ l = \{1, 2, \ldots, p\}.$$

1. Если $i = j$, то

$$1 = (e'_i, e'_i) = ((\varphi_{1i}e_1 + \varphi_{2i}e_2 + \cdots + \varphi_{pi}e_p), (\varphi_{1i}e_1 + \varphi_{2i}e_2 + \cdots + \varphi_{pi}e_p)) =$$

$$= \left(\sum_{j=1}^{p} \varphi_{ji}e_j,\ \sum_{k=1}^{p} \varphi_{ki}e_k\right) = \sum_{j,k=1}^{p} \varphi_{ji}\varphi_{ki}(e_j, e_k) = \varphi_{1i}^2 + \varphi_{2i}^2 + \ldots + \varphi_{pi}^2,$$

так как скалярное произведение (e_j, e_k) не равно нулю только при $j = k$. Отсюда следует, что столбцы матрицы Φ имеют единичную длину.
2. Если $i \neq j$, то

$$0 = (e'_i, e'_j) = ((\varphi_{1i}e_1 + \varphi_{2i}e_2 + \cdots + \varphi_{pi}e_p), (\varphi_{1j}e_1 + \varphi_{2j}e_2 + \cdots + \varphi_{pj}e_p)) =$$

$$= \left(\sum_{j=1}^{p} \varphi_{ji}e_j,\ \sum_{k=1}^{p} \varphi_{ki}e_k\right) = \sum_{j,k=1}^{p} \varphi_{ji}\varphi_{ki}(e_j, e_k) = \varphi_{1i}\varphi_{1j} + \varphi_{2i}\varphi_{2j} + \varphi_{pi}\varphi_{pj},$$

а значит столбцы матрицы Φ попарно ортогональны.
III. Рассмотрим произведение $\Phi^T\Phi$. В силу того, что столбцы матрицы Φ (строки матрицы Φ^T) попарно ортогональны и имеют единичную длину, получим

$$\Phi^T\Phi = \begin{pmatrix} \varphi_{11} & \varphi_{21} & \cdots & \varphi_{p1} \\ \varphi_{12} & \varphi_{22} & \cdots & \varphi_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{1p} & \varphi_{2p} & \cdots & \varphi_{pp} \end{pmatrix} \begin{pmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1p} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{p1} & \varphi_{p2} & \cdots & \varphi_{pp} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} = E.$$

Таким образом,

$$\Phi^T\Phi = E,$$

а значит, матрица Φ^T является левой обратной для Φ. По свойствам квадратных матриц, Φ^T является и правой обратной для Φ, откуда

$$\Phi\Phi^T = E,$$

а значит

$$\Phi^{-1} = \Phi^T.$$

II. В силу того, что $\Phi^T = \Phi^{-1}$ и свойств обратной матрицы, получим

$$\Phi^T\Phi = \Phi^{-1}\Phi = \Phi\Phi^{-1} = \Phi\Phi^T = E.$$

Последнее равенство можно записать в виде

$$\begin{pmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1p} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{p1} & \varphi_{p2} & \cdots & \varphi_{pp} \end{pmatrix} \begin{pmatrix} \varphi_{11} & \varphi_{21} & \cdots & \varphi_{p1} \\ \varphi_{12} & \varphi_{22} & \cdots & \varphi_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{1p} & \varphi_{2p} & \cdots & \varphi_{pp} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix},$$

из которого (по аналогичным рассуждениям пункта I) следует, что строки матрицы Φ попарно ортогональны и имеют единичную длину. □

Для полноты картины введем и боле общее определение.

Определение 4.1.3 *Матрица* Φ*, для которой выполнено* $\Phi^{-1} = \Phi^T$*, называется ортогональной.*

Оказывается, что любая ортогональная матрица переводит ортонормированный базис в ортонормированный, а значит для любой ортогональной матрицы справедливы пункты I и II предыдущей теоремы.

4.2 Обратно в МГК

Давайте теперь применим изученный аппарат к исследованию МГК. Пусть F – матрица исходных данных размера $[n \times p]$, состоящая из n объектов, каждый из которых обладает p признаками:

$$F = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

Будем также считать, что центрирование уже произведено. Наша цель – перейти от одного ортонормированного базиса (исходного) к другому ортонормированному (который дает МГК). В таком случае матрица перехода Φ будет ортогональной, а значит, как мы только что отметили, для матрицы

$$\Phi = \begin{pmatrix} \varphi_{11} & \varphi_{12} & \dots & \varphi_{1p} \\ \varphi_{21} & \varphi_{22} & \dots & \varphi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{p1} & \varphi_{p2} & \dots & \varphi_{pp} \end{pmatrix}$$

справедливы следующие утверждения:

1. $\Phi^T = \Phi^{-1}$.
2. Строки и столбцы матрицы Φ – ортонормированные векторы.

Строки матрицы F, как обычно, представляют собой объекты (или векторы) исходного набора данных. В дальнейших же преобразованиях нам будет удобнее, чтобы объект был не строкой, а столбцом, поэтому пусть столбец

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

– это какой-то объект X (какая-то строка матрицы F), координаты которого заданы в исходном базисе (будем считать его новым), а столбец

$$Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix}$$

– это тот же объект, координаты которого заданы в другом базисе (скажем, старом), и пусть Φ – матрица перехода.

Из соотношения $Z = \Phi X$ выразим вектор X:

$$X = \Phi^{-1} Z = \Phi^T Z,$$

где

$$\Phi^T = \begin{pmatrix} \varphi_{11} & \varphi_{21} & \cdots & \varphi_{p1} \\ \varphi_{12} & \varphi_{22} & \cdots & \varphi_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{1p} & \varphi_{2p} & \cdots & \varphi_{pp} \end{pmatrix}.$$

Выражение $X = \Phi^T Z$ в матричных обозначениях записывается, как

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} \varphi_{11} & \varphi_{21} & \cdots & \varphi_{p1} \\ \varphi_{12} & \varphi_{22} & \cdots & \varphi_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{1p} & \varphi_{2p} & \cdots & \varphi_{pp} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix}$$

Выполняя умножение матриц Φ^T и Z, получим:

$$x_1 = \varphi_{11} z_1 + \varphi_{21} z_2 + \cdots + \varphi_{p1} z_p$$

$$x_2 = \varphi_{12} z_1 + \varphi_{22} z_2 + \cdots + \varphi_{p2} z_p$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$x_p = \varphi_{1p} z_1 + \varphi_{2p} z_2 + \cdots + \varphi_{pp} z_p.$$

Так как z_i – число, то вектор X может быть представлен в следующем виде:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} \varphi_{11} \\ \varphi_{12} \\ \vdots \\ \varphi_{1p} \end{pmatrix} z_1 + \begin{pmatrix} \varphi_{21} \\ \varphi_{22} \\ \vdots \\ \varphi_{2p} \end{pmatrix} z_2 + \cdots + \begin{pmatrix} \varphi_{p1} \\ \varphi_{p2} \\ \vdots \\ \varphi_{pp} \end{pmatrix} z_p.$$

В результате можно ввести следующее определение.

Определение 4.2.1 *Представление вектора X в виде*

$$X = \begin{pmatrix} \varphi_{11} \\ \varphi_{12} \\ \vdots \\ \varphi_{1p} \end{pmatrix} z_1 + \begin{pmatrix} \varphi_{21} \\ \varphi_{22} \\ \vdots \\ \varphi_{2p} \end{pmatrix} z_2 + \cdots + \begin{pmatrix} \varphi_{p1} \\ \varphi_{p2} \\ \vdots \\ \varphi_{pp} \end{pmatrix} z_p,$$

где $(\varphi_{i1}\ \varphi_{i2}\ \ldots\ \varphi_{ip})$ – строки матрицы перехода, а z_i – соответствующие координаты объекта в старом базисе, называется разложением вектора X по его главным компонентам.

Обозначим

$$\varphi_1 = \begin{pmatrix} \varphi_{11} \\ \varphi_{12} \\ \vdots \\ \varphi_{1p} \end{pmatrix},\ \varphi_2 = \begin{pmatrix} \varphi_{21} \\ \varphi_{22} \\ \vdots \\ \varphi_{2p} \end{pmatrix},\ \ldots,\ \varphi_p = \begin{pmatrix} \varphi_{p1} \\ \varphi_{p2} \\ \vdots \\ \varphi_{pp} \end{pmatrix}$$

– столбцы матрицы Φ^T (строки матрицы Φ). Тогда разложение вектора X по главным компонентам можно представить в следующем виде:

$$X = \sum_{i=1}^{p} \varphi_i z_i.$$

Предположим, мы хотим оставить $k \leq p$ первых главных компонент:

$$\widehat{X} = \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \vdots \\ \widehat{x}_p \end{pmatrix} = \begin{pmatrix} \varphi_{11} \\ \varphi_{12} \\ \vdots \\ \varphi_{1p} \end{pmatrix} \cdot z_1 + \begin{pmatrix} \varphi_{21} \\ \varphi_{22} \\ \vdots \\ \varphi_{2p} \end{pmatrix} \cdot z_2 + \cdots + \begin{pmatrix} \varphi_{k1} \\ \varphi_{k2} \\ \vdots \\ \varphi_{kp} \end{pmatrix} \cdot z_k,$$

или, в более компактном виде:

$$\widehat{X} = \sum_{i=1}^{k} \varphi_i z_i.$$

Логично ожидать, что чем меньше главных компонент мы возьмем, тем больше информации об исходных объектах будет потеряно. Так как X – это некоторый объект матрицы исходных данных, обладающий p признаками, а каждый признак, в свою очередь, является случайной величиной, то и объект также является случайной величиной. Поэтому можно рассмотреть так называемую ошибку MSE – ошибку в среднеквадратичном, равную математическому ожиданию квадрата нормы разности X и $\widehat{X}$, то есть

$$\mathsf{E}\left\|X - \widehat{X}\right\|^2 = \mathsf{E}\left\|\sum_{i=1}^{p} \varphi_i z_i - \sum_{i=1}^{k} \varphi_i z_i\right\|^2 =$$

$$= \mathsf{E}\left\|\sum_{i=k+1}^{p} \varphi_i z_i\right\|^2 = \mathsf{E}\left(\sum_{i=k+1}^{p} \varphi_i z_i, \sum_{i=k+1}^{p} \varphi_i z_i\right).$$

Учитывая то, что векторы φ_i попарно ортогональны, а длина каждого равна единице, то есть

$$(\varphi_i, \varphi_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases},$$

где $i, j = \{k+1, \ldots, p\}$, и учитывая свойство линейности скалярного произведения получим, что

$$\mathsf{E}\left\|X - \widehat{X}\right\|^2 = \mathsf{E}\left(\sum_{i=k+1}^{p} \varphi_i z_i, \sum_{i=k+1}^{p} \varphi_i z_i\right) = \mathsf{E}\left(\sum_{i=k+1}^{p} z_i^2\right) = \sum_{i=k+1}^{p} \mathsf{E} z_i^2.$$

Так как в матричном виде выражение $Z = \Phi X$, переписывается, как

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix} = \begin{pmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1p} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{p1} & \varphi_{p2} & \cdots & \varphi_{pp} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix},$$

то каждая координата z_i равна произведению i-ой строки матрицы Φ на столбец X, а учитывая, что φ_i – это i-ый столбец матрицы Φ^T, получим:

$$z_i = \varphi_i^T X.$$

С другой стороны, z_i можно представить в виде произведения транспонированного вектора X на вектор φ_i:

$$z_i = X^T \varphi_i.$$

Возвращаясь к рассматриваемому выражению и учитывая свойства математического ожидания, получим

$$\mathsf{E}\left\|X - \widehat{X}\right\|^2 = \sum_{i=k+1}^{p} \mathsf{E} z_i^2 = \sum_{i=k+1}^{p} \mathsf{E}\left(z_i \cdot z_i\right) =$$

$$= \sum_{i=k+1}^{p} \mathsf{E}\left(\varphi_i^T X \cdot X^T \varphi_i\right) = \sum_{i=k+1}^{p} \varphi_i^T \mathsf{E}\left(X \cdot X^T\right) \varphi_i.$$

Обозначим $\Theta = \mathsf{E}\left(X \cdot X^T\right)$.

Замечание 4.2.1 *Легко заметить, что Θ – это матрица ковариации для центрированного случайного вектора X.*

В итоге приходим к выражению, которое и будем стремиться минимизировать:

$$\mathsf{E}\left\|X-\widehat{X}\right\|^2=\sum_{i=k+1}^{p}\varphi_i^T\Theta\varphi_i\longrightarrow\min_{\varphi_i}.$$

Напомним определения, которые мы будем использовать в дальнейшем.

Определение 4.2.2 *Ненулевой вектор φ называют собственным вектором матрицы Θ, если для некоторого числа λ выполняется соотношение*

$$\Theta\varphi=\lambda\varphi.$$

При этом число λ называют собственным числом (или собственным значением) матрицы Θ.

Замечание 4.2.2 *В качестве замечания отметим, что матрица ковариаций является симметричной, то есть $\Theta=\Theta^T$, и имеет p неотрицательных собственных чисел с учетом кратности. Кроме того, из ее собственных векторов можно построить ортонормированный набор из p элементов.*

Теорема 4.2.1 *Пусть*

$$\lambda_1\geq\lambda_2\geq...\geq\lambda_p\geq0$$

– собственные числа матрицы ковариаций. Минимум выражения

$$\mathsf{E}\left\|X-\widehat{X}\right\|^2=\sum_{i=k+1}^{p}\varphi_i^T\Theta\varphi_i$$

достигается в случае, если φ_i – ортонормированные собственные векторы, соответствующие наименьшим собственным числам λ_i матрицы Θ, причем на них

$$\min_{\varphi_i}\mathsf{E}\left\|X-\widehat{X}\right\|^2=\min_{\varphi_i}\sum_{i=k+1}^{p}\varphi_i^T\Theta\varphi_i=\sum_{i=k+1}^{p}\lambda_i.$$

Доказательство. Для доказательства теоремы нам необходимо найти такие векторы φ_i, при которых значение этого выражения было бы минимальным. Иными словами, перед нами задача поиска условного экстремума. Применим для ее решения метод множителей Лагранжа, где $\mathsf{E}\left\|X-\widehat{X}\right\|^2$ – минимизируемое выражение, $|\varphi_i|=1,\varphi_i^T\varphi_i=1$ – ограничения. Напомним, что метод

Лагранжа заключается в том, что поиск условного экстремума сводится к поиску экстремума так называемой функции Лагранжа. Функция Лагранжа имеет вид:

$$L(\varphi, \lambda) = \sum_{i=k+1}^{p} \varphi_i^T \Theta \varphi_i - \sum_{i=k+1}^{p} \lambda_i \left(\varphi_i^T \varphi_i - 1\right).$$

Перепишем это выражение, записав слагаемые в правой части под одним знаком суммы:

$$L(\varphi, \lambda) = \sum_{i=k+1}^{p} \left(\varphi_i^T \Theta \varphi_i - \lambda_i \left(\varphi_i^T \varphi_i - 1\right)\right).$$

Продифференцируем функцию Лагранжа по всем переменным, приравняем частные производные к нулю и решим полученную систему уравнений. В нашем случае неизвестные – это как векторы φ_i, так и числа λ_i. Можно показать, что

$$\frac{\partial \left(\varphi_i^T \Theta \varphi_i\right)}{\partial \varphi_i} = (\Theta + \Theta^T)\varphi_i.$$

В случае симметричной матрицы Θ (коей является матрица ковариаций), получаем

$$\frac{\partial \left(\varphi_i^T \Theta \varphi_i\right)}{\partial \varphi_i} = 2\Theta\varphi_i,$$

откуда

$$\frac{\partial L(\varphi, \lambda)}{\partial \varphi_i} = 2\Theta\varphi_i - 2\lambda_i\varphi_i = 0.$$

Иначе последнее выражение переписывается в виде

$$\Theta\varphi_i = \lambda_i\varphi_i.$$

Написанное выше – не что иное, как определение собственных чисел и собственных векторов матрицы Θ. Если мы подставим это выражение в L, то учитывая, что $\sum_{i=k+1}^{p} \lambda_i \left(\varphi_i^T \varphi_i - 1\right) = 0$ (в силу уравнений связи), получим:

$$\mathsf{E}\left\|X - \widehat{X}\right\|^2 = \sum_{i=k+1}^{p} \varphi_i^T \lambda_i \varphi_i = \sum_{i=k+1}^{p} \lambda_i.$$

□

Иными словами, выражение будет минимальным в случае, когда λ_i – наименьшие из собственных чисел матрицы Θ. Таким образом получается, что

при сокращении размерности, чтобы потерять как можно меньше информации, разумно в качестве вектора весов первой ГК брать собственный вектор, отвечающий наибольшему из собственных чисел матрицы Θ, в качестве вектора весов второй ГК брать собственный вектор, отвечающий следующему по величине собственному числу матрицы Θ и так далее.

Замечание 4.2.3 *Отметим без дополнительных пояснений, что i-ое собственное число λ_i равно дисперсии счётов i-ой ГК.*

4.3 Выбор количества ГК

Важным остается вопрос, а сколько главных компонент достаточно оставлять, чтобы не потерять слишком много информации об исходных объектах? Введем следующее определение.

Определение 4.3.1 *Пусть*

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$$

– собственные числа матрицы ковариаций, а в качестве векторов весов первых k ГК взяты собственные векторы матрицы ковариации, отвечающие наибольшим собственным значениям $\lambda_1, \lambda_2, \ldots, \lambda_k$. Величина

$$\delta_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

называется долей объясненной дисперсии.

Замечание 4.3.1 *Величина $1 - \delta_k$ называется долей необъясненной (остаточной) дисперсии.*

Нетрудно заметить, что δ_k принимает значения от нуля до единицы и показывает, какая часть дисперсии учитывается при использовании первых k ГК относительно всей дисперсии. Иными словами, чем ближе δ_k к единице, тем меньше информации об исходных объектах мы теряем.

Имеет смысл оставлять столько главных компонент, чтобы добавление последующих не влекло существенного изменения доли объясненной дисперсии. Поясним это на рисунке 8.

По оси абсцисс указано количество оставляемых главных компонент, а по оси ординат – соответствующая доля объясненной дисперсии. Можно заметить, что, начиная с 3-ей ГК доля объясненной дисперсии изменяется незначительно, при этом три главных компоненты описывают около 95% всей дисперсии. В такой ситуации имеет смысл оставлять именно три главных компоненты.

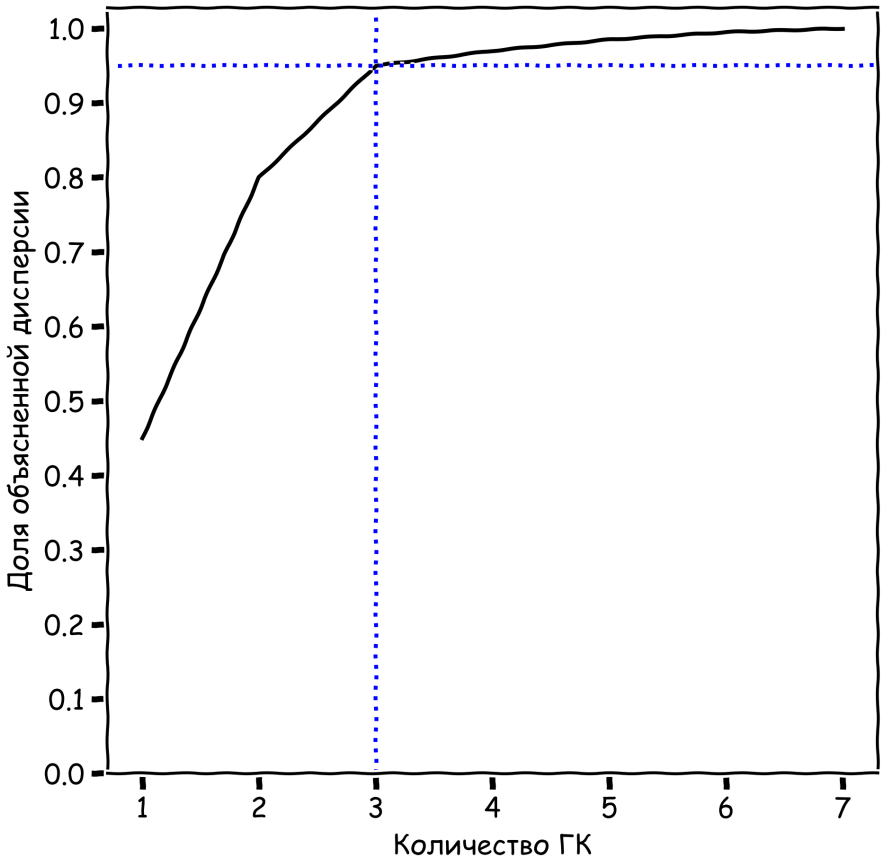

Рис. 9: Выбор количества ГК.

4.4 Алгоритм

Рассмотрим последовательность действий для нахождения главных компонент. Пусть F – центрированная матрица, содержащая информацию об n объектах, обладающих p признаками.

$$F = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

Учитывая, что теперь, во-первых, мы имеем дело с выборкой, а не со случайной величиной, а во-вторых объекты представляют собой строки в матрице исходных данных, разумно выполнить следующие шаги:

1. Найти выборочную ковариационную матрицу из соотношения

$$\Theta = \frac{1}{n}\left(F^T \cdot F\right).$$

2. Найти собственные числа λ_i матрицы Θ, $i = \{1, 2, \dots, p\}$.

3. Найти ортонормированные собственные векторы φ_i матрицы Θ, соответствующие собственным числам λ_i.

4. Выбрать необходимое число главных компонент. При этом в качестве вектора весов первой главной компоненты разумно, чтобы MSE была наименьшей, взять собственный вектор матрицы Θ, соответствующий наибольшему собственному числу этой матрицы. В качестве вектора весов второй ГК – собственный вектор, соответствующий второму по величине собственному числу и так далее.

5. Найти новые координаты (или счёты) объектов в выбранном базисе, выполнив умножение

$$Z = F\Phi,$$

учитывая, что координаты выбранных собственных векторов являются столбцами матрицы Φ, причем в первом столбце Φ стоят координаты вектора весов, отвечающего наибольшему собственному числу матрицы Θ, во втором – координаты вектора весов, отвечающего второму по величине собственному числу, и так далее.

Кроме того, в качестве замечания отметим, что значения собственных чисел λ_i для выбранных ГК равны выборочным дисперсиям i-ых счётов.

4.5 Восстановление признаков по главным компонентам

Метод главных компонент используется для сокращения количества признаков, однако часто требуется решить обратную задачу, то есть по главным компонентам восстановить начальные признаки объектов (естественно, если количество ГК меньше, чем размерность исходного пространства объектов, то с некоторой ошибкой, так как при уменьшении размерности информация все-таки теряется). Такая необходимость может возникнуть, например, если в результате построения главных компонент обнаружен выброс. Тогда, восстановив начальные признаки, можно более внимательно рассмотреть этот объект и проанализировать «почти исходные» признаки на предмет несоответсвия тенденции.

Итак, как же происходит восстановление? Пусть, как и ранее, Φ – матрица, столбцы которой отвечают координатам нормированных собственных векторов – векторов весов. Тогда

$$Z_{[n\times p]} = F_{[n\times p]}\Phi_{[p\times p]},$$

и матрица счетов имеет размерность $[n \times p]$. Тогда старые центрированные координаты восстанавливаются без всяких потерь домножением всего равен-

ства справа на Φ^T, откуда

$$Z\Phi^T = F\Phi\Phi^T = FE = F,$$

так как, в силу ортогональности Φ, $\Phi\Phi^T = E$ – единичная матрица. Однако обычно количество ГК, которые мы оставляем, меньше, чем размерность исходного пространства. Оставив их k штук, получим матрицу Φ размера $[p \times k]$ и вектор счётов

$$Z_{[n\times k]} = F_{[n\times p]}\Phi_{[p\times k]}$$

размера $[n \times k]$. Домножим справа на Φ^T и заметим, что теперь, хотя произведение $\Phi\Phi^T$ и будет иметь размер $[p \times p]$, оно не будет единичной матрицей. Поэтому

$$Z\Phi^T = F\Phi\Phi^T = \widetilde{F},$$

где $\widetilde{F}$ – матрица с координатами приближенно восстановленных центрированных исходных объектов.

То, что восстановление оказывается приближенным, понятно: ведь при уменьшении размерности пространства, как не раз говорилось, теряется информация. Это же можно пояснить и наглядно, посмотрев на рисунок 9. Сопоставив объекту x_1, имеющему два признака (x_{11}, x_{12}), лишь один, восстановление старых признаков, опираясь только на знание о первой главной компоненте, происходит с ошибкой в признаки $(\widetilde{x}_{11}, \widetilde{x}_{12})$.

На самом деле, для восстановления исходных признаков, к полученным после восстановления нужно добавить еще ранее вычтенные средние $\overline{X'_j}$, $j \in \{1, 2, ..., p\}$ (ведь у исходных даных не предполагалось центрированности). Введем матрицу $\overline{X}$ размера $[n \times p]$ следующим образом

$$\overline{X} = \begin{pmatrix} \overline{X'_1} & \overline{X'_2} & \dots & \overline{X'_p} \\ \overline{X'_1} & \overline{X'_2} & \dots & \overline{X'_p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{X'_1} & \overline{X'_2} & \dots & \overline{X'_p} \end{pmatrix}.$$

Тогда окончательное восстановление (или приближенное восстановление) по счётам Z может быть записано, как

$$\widetilde{F}' = Z\Phi^T + \overline{X},$$

причем, если Φ имеет размер $[p \times p]$, то $\widetilde{F}' = F'$, где F' – матрица исходных объектов до центрирования.

4.6 Еще один взгляд на пример

Покажем на уже знакомом примере с продавцами автомобилей, как находить главные компоненты при помощи матрицы ковариаций. А также выясним, какую часть информации мы потеряли, если использовали всего одну

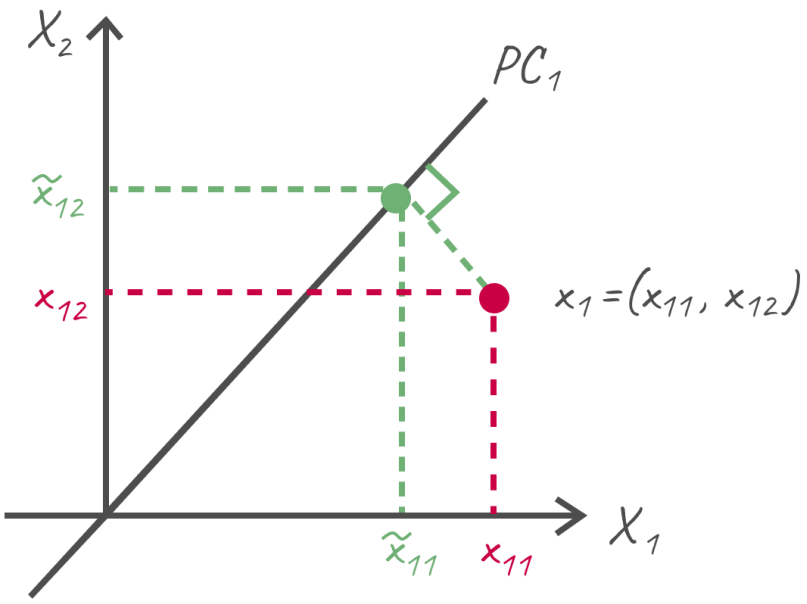

Рис. 10: Геометрия восстановления

главную компоненту, и на что будут похожи восстановленные данные. Напомним условия задачи.

Таблица исходных данных.

Сотрудник (x_i)	**Премиум** (X_1)**, шт**	**Эконом** (X_2)**, шт**
1	9	19
2	6	22
3	11	27
4	12	25
5	7	22

Средние значения признаков:

$$\overline{X_1'} = 9,$$

$$\overline{X_2'} = 23$$

Таблица центрированных исходных данных.

Сотрудник (x_i)	**Признак** X_1	**Признак** X_2
1	0	-4
2	-3	-1
3	2	4
4	3	2
5	-2	-1

Составим матрицу F:

$$F = \begin{pmatrix} 0 & -4 \\ -3 & -1 \\ 2 & 4 \\ 3 & 2 \\ -2 & -1 \end{pmatrix}$$

и выборочную ковариационную матрицу Θ:

$$\Theta = \frac{1}{n}\left(F^T \cdot F\right) = \frac{1}{5}\begin{pmatrix} 0 & -3 & 2 & 3 & -2 \\ -4 & -1 & 4 & 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -4 \\ -3 & -1 \\ 2 & 4 \\ 3 & 2 \\ -2 & -1 \end{pmatrix}.$$

В результате умножения матриц получим:

$$\Theta = \frac{1}{5}\begin{pmatrix} 26 & 19 \\ 19 & 38 \end{pmatrix}.$$

Найдем собственные числа этой матрицы. Напомним, что собственные числа находятся из условия, что определитель разности Θ и λE равен нулю:

$$\left|\Theta - \lambda E\right| = 0,$$

где E – единичная матрица. В нашем случае приходим к уравнению

$$\begin{vmatrix} \frac{26}{5} - \lambda & \frac{19}{5} \\ \frac{19}{5} & \frac{38}{5} - \lambda \end{vmatrix} = 0.$$

По правилу нахождения определителя второго порядка, получим:

$$\left(\frac{26}{5} - \lambda\right)\left(\frac{38}{5} - \lambda\right) - \left(\frac{19}{5}\right)^2 = 0.$$

Решая это уравнение как квадратное, найдем λ_1 и λ_2:

$$\lambda_1 = \frac{32 + \sqrt{397}}{5},$$

$$\lambda_2 = \frac{32 - \sqrt{397}}{5}.$$

Так как $\lambda_1 = \max(\lambda_1, \lambda_2)$, то соответствующий нормированный собственный вектор будет вектором весов первой главной компоненты. Давайте его найдем. Так как по определению:

$$\Theta\varphi_i = \lambda_i\varphi_i,$$

получим:

$$\begin{pmatrix} 26 & 19 \\ 19 & 38 \end{pmatrix} \cdot \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \end{pmatrix} = (32 + \sqrt{397}) \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \end{pmatrix}.$$

Тогда для нахождения φ_1 решим соответствующую систему уравнений:

$$\begin{cases} (-6 - \sqrt{397})\varphi_{11} + 19\varphi_{21} = 0 \\ 19\varphi_{11} + (6 - \sqrt{397})\varphi_{21} = 0 \end{cases}.$$

Система имеет бесконечное множество решений. Возьмем φ_{11} в качестве базисной переменной, φ_{21} в качестве свободной. Тогда при $\varphi_{21} = 1$, получим $\varphi_{11} \approx 0.733$. Напомним, что длина вектора весов главной компоненты должна равняться единице, поэтому выполним нормирование, а именно, поделим каждую координату на длину вектора:

$$\varphi_{11} \approx \frac{0.733}{\sqrt{1^2 + 0.733^2}} \approx 0.591,$$

$$\varphi_{21} \approx \frac{1}{\sqrt{1^2 + 0.733^2}} \approx 0.807.$$

Тогда, если $Z_1 = F\varphi_1$, то получим следующее выражение для счётов первой ГК:

$$Z_1 = \begin{pmatrix} 0 & -4 \\ -3 & -1 \\ 2 & 4 \\ 3 & 2 \\ -2 & -1 \end{pmatrix} \cdot \begin{pmatrix} 0.591 \\ 0.807 \end{pmatrix} = \begin{pmatrix} -3.226 \\ -2.580 \\ 4.409 \\ 3.387 \\ -1.989 \end{pmatrix}.$$

Найдем долю объясненной дисперсии в случае, если мы оставляем одну главную компоненту.

$$\delta_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{32 + \sqrt{397}}{(32 - \sqrt{397}) + (32 + \sqrt{397})} \approx 0.811.$$

Попробуем выполнить обратную задачу, а именно восстановить начальные признаки, по первой главной компоненте. Тогда

$$\widetilde{F}' = Z\Phi^T + \overline{X} =$$

$$= \begin{pmatrix} -3.226 \\ -2.580 \\ 4.409 \\ 3.387 \\ -1.989 \end{pmatrix} \cdot \begin{pmatrix} 0.591 & 0.807 \end{pmatrix} + \begin{pmatrix} 9 & 23 \\ 9 & 23 \\ 9 & 23 \\ 9 & 23 \\ 9 & 23 \end{pmatrix} = \begin{pmatrix} 7.093 & 20.397 \\ 7.475 & 20.918 \\ 11.606 & 26.558 \\ 11.002 & 25.733 \\ 7.825 & 21.395 \end{pmatrix}.$$

Можно заметить, что если округлить восстановленные по первой ГК данные до целых, то они весьма похожи на исходные:

$$F' = \begin{pmatrix} 9 & 19 \\ 6 & 22 \\ 11 & 27 \\ 12 & 25 \\ 7 & 22 \end{pmatrix}, \quad \widetilde{F}' = \begin{pmatrix} 7 & 20 \\ 7 & 21 \\ 12 & 27 \\ 11 & 26 \\ 8 & 25 \end{pmatrix},$$

где F' – исходные данные, $\widetilde{F}'$ – восстановленные.

5 Примеры использования МГК

5.1 Пример визуализации

Рассмотрим пример использования МГК для определения выбросов и возможности разделения объектов на группы. В качестве исходных данных возьмем датасет, содержащий набор изображений букв латинского алфавита[2]. Данные получены в результате случайного искажения пиксельных изображений 26 заглавных букв из 20 различных коммерческих шрифтов. Пример представлен на рисунке 10.

Рис. 11: Пример искаженных букв.

[2]https://archive.ics.uci.edu/ml/datasets/Letter+Recognition

Каждый объект имеет 16 признаков. Изначально признаки описывали статистические характеристики распределения пикселей (горизонтальная и вертикальные координаты центра наименьшего прямоугольника, содержащего в себе все «закрашенные» пиксели, ширина и высота этого прямоугольника, общее количество «закрашенных» пикселей и так далее). Далее авторы датасета масштабировали признаки так, чтобы они могли принимать целые значения из диапазона от 0 до 15.

Для демонстрации идеи выберем из всего набора данных только объекты, соответствующие буквам A, B и C и визуализируем данные, используя 2 главные компоненты. Результаты представлены на рисунке 11.

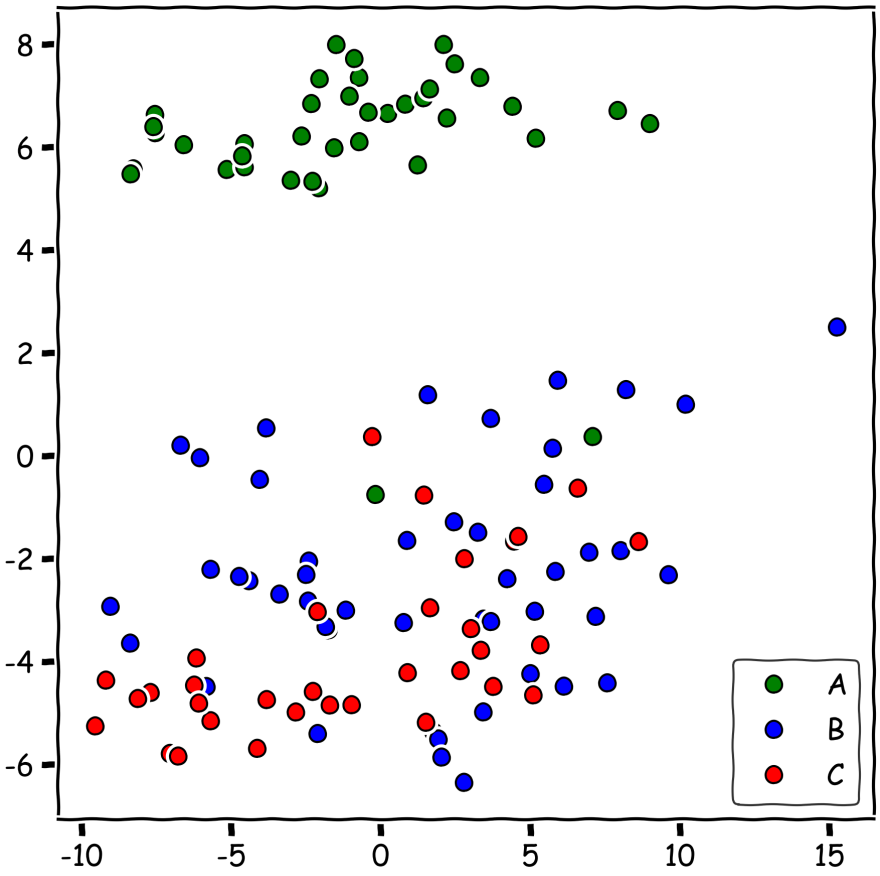

Рис. 12: Визуализация по первым двум ГК.

Взглянув на полученное изображение, можно заметить, что объекты достаточно неплохо делятся на группы, особенно хорошо это прослеживается для буквы A. При этом имеют место очевидные выбросы: прекрасно видны две зеленые точки, отвечающие буквам A, затесавшиеся среди синих и красных, отвечающих B и C.

Возникает вполне резонный вопрос: а сколько информации мы сохраним, если оставим только две ГК? Построим график зависимости доли объясненной дисперсии от количества главных компонент (рисунок 12).

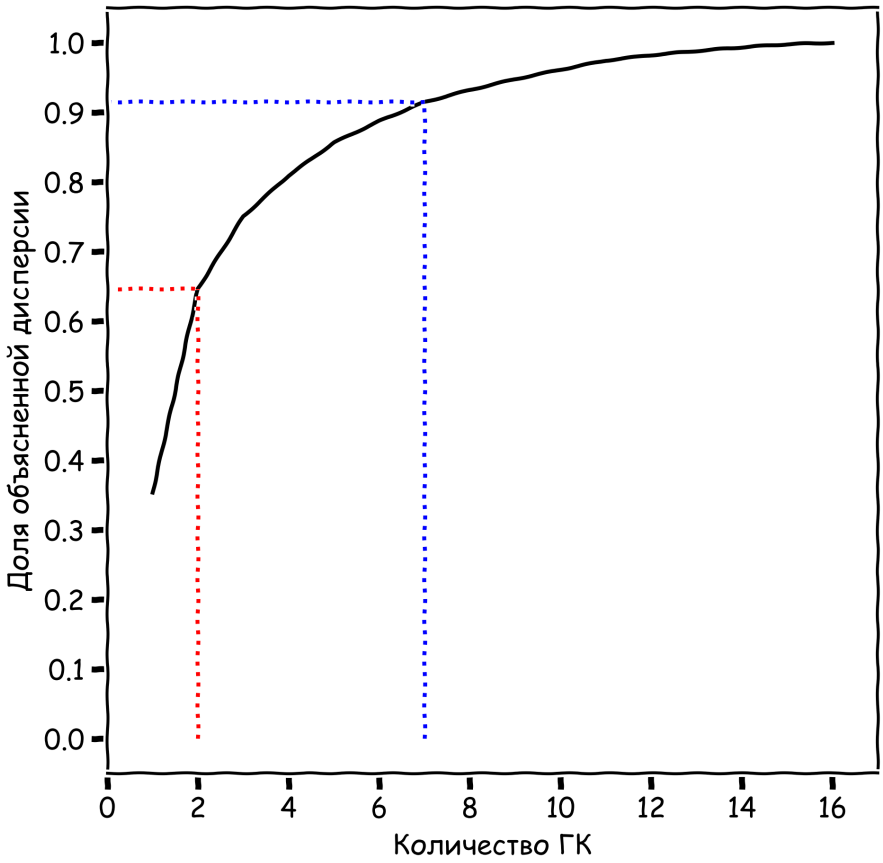

Рис. 13: Выбор количества ГК.

Можно заметить, что если оставлять всего лишь две главных компоненты, то мы сохраним порядка 65% информации, при этом если увеличить количество ГК до семи, то доля объясненной дисперсии составит уже более 0.9, то есть сохранится более 90% информации.

5.2 Пример компрессии изображений

В качестве еще одного примера использования МГК можно рассмотреть компрессию изображений: для уменьшения объемов данных используют МГК, а затем восстанавливают изображение, но, естественно, с некоторыми потерями.

Опробуем метод на достаточно известном датасете, содержащем изображения рукописных цифр[3]. Изображения можно разделить на 10 классов, где класс – это цифра. Каждое изображение имеет размер 8×8, что соответствует 64 атрибутам. Каждый атрибут принимает целые значения в диапазоне от 0 до 16 (градации серого от белого до черного). Пример изображений представлен на рисунке 13.

Заметим, что исходные данные сами по себе не самого высокого качества.

[3]https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

Рис. 14: Исходные изображения

Значит при восстановлении картина будет еще хуже. Определимся сначала с тем, какое количество ГК нам интересно будет рассматривать для дальнейшего восстановления исходных данных. Построим график зависимости доли объясненной дисперсии от количества главных компонент. Он представлен на рисунке 14.

Можно заметить, что при использовании менее десяти главных компонент мы потеряем более четверти исходной информации, в то время как оставлять более сорока ГК не видится целесообразным. Восстановим данные для случаев 2, 5, 10, 20 и 40 ГК. Они представлены на рисунке 15. Слева указано количество использованных главных компонент (k) и доля объясненной дисперсии (δ). При использовании 64 главных компонент изображения восстанавливаются без потерь и совпадают с оригиналом. Как мы и предполагали, основываясь на графике зависимости доли объясненной дисперсии от количества используемых ГК, для данного случая вполне достаточно использовать 20 ГК для адекватного восстановления исходных данных. При использовании же 40 ГК, восстановленные данные практически ничем не отличаются от оригинала.

6 МГК и дисперсия

Метод главных компонент из-за работы с выборочной дисперсией является чувствительным к единицам измерения исходных данных. Для наглядности рассмотрим такой пример: пусть в качестве объектов выступают автомобильные дороги. Пусть они обладают двумя признаками: протяженность (в метрах) и среднее количество ДТП в год.

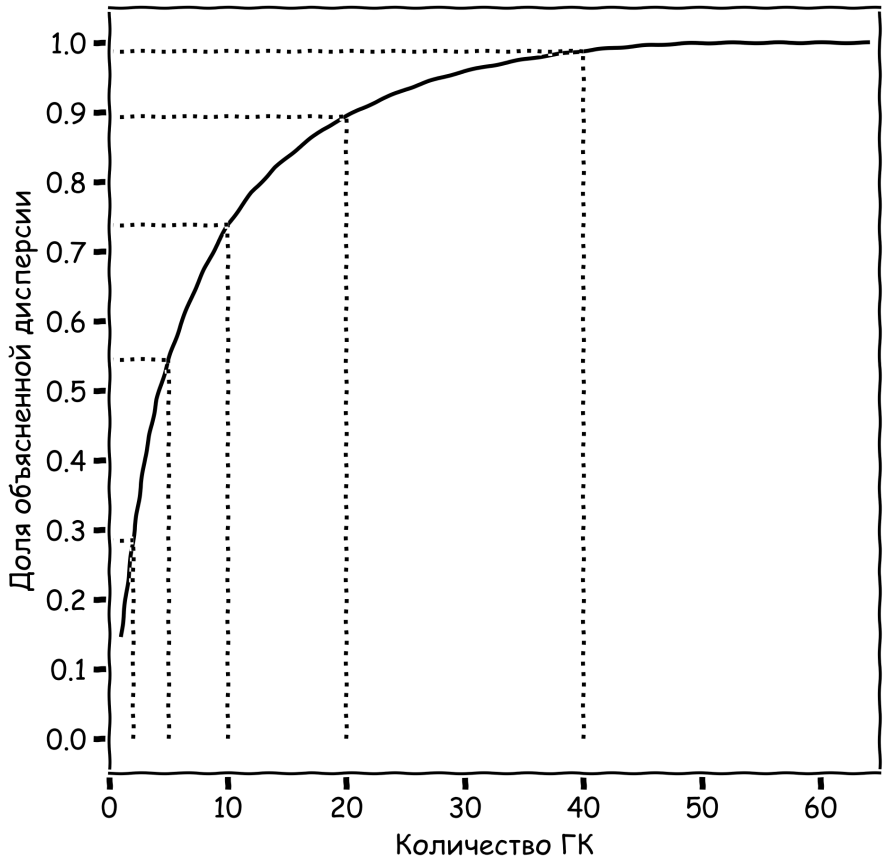

Рис. 15: Выбор количества ГК.

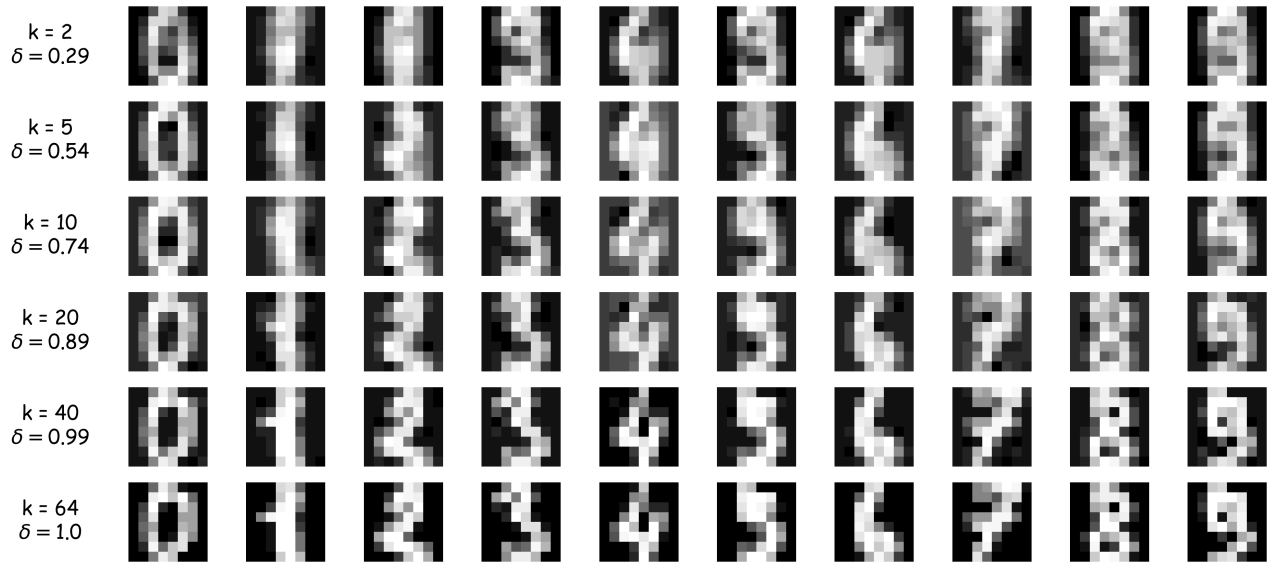

Рис. 16: Восстановление исходных данных

	E95	**M4**	**M10**
Протяженность, м	37700000	1517000	697000
Среднее количество ДТП, шт	345	84	51

При использовании евклидова расстояния разница в один метр по первой координате будет вносить тот же вклад, что и разница в одно ДТП по второй, что не совсем корректно. Напомним, что в процессе поиска главных компонент осуществляется поиск направления наибольшего разброса.

$$S^2_{\text{Протяженность}} \approx 2.5 \cdot 10^{12},$$

$$S^2_{\text{ДТП}} = 25941.$$

Очевидно, что при нахождении первой главной компоненты, выборочная дисперсия для ДТП будет играть очень маленькую роль, и направление ГК будет мало отличаться от направления оси, отвечающей за протяженность. Хотя среднее количество ДТП является тоже очень важной характеристикой, по которой исходные объекты, как видно из таблицы, серьезно отличаются. Для устранения этой проблемы возможно использовать какую-либо нормировку, чтобы масштабы признаков были сравнимы. Например, если использовать линейную нормировку, получим следующие значения:

	E95	**M4**	**M10**
Протяженность	1	0.27	0
Среднее количество ДТП	1	0.11	0

Теперь выборочная дисперсия в каждом случае будет сопоставима:

$$S^2_{\text{Протяженность}} \approx 0.27,$$

$$S^2_{\text{ДТП}} \approx 0.3.$$

7 Заключение

Резюмируя, можно сказать, что метод главных компонент является достаточно эффективным способом уменьшения размерности исходного пространства признаков. Наиболее часто он используется для визуализации исходного набора данных (строят две или три первых главных компоненты для того, чтобы можно было визуализировать исходные данные в виде объектов на плоскости или в пространстве). Кроме того, МГК имеет достаточно большое количество приложений в различных областях: компрессия изображений, подавление шума на изображениях, биоинформатика и многое другое. Важно учитывать, что признаки могут иметь как разную область изменения, так и представлять различные характеристики объектов, поэтому для более корректной работы метода обязательно нужно производить предварительную нормировку исходных данных.