Введение в многомерный статистический анализ
2021-2022 учебный год

Автор лекции: Тамбовцева А.А.

Лекция 1. Метод главных компонент: постановка задачи и геометрическая интерпретация

1.1 Введение в метод главных компонент

Постановка задачи

- На входе: p-мерный массив, p переменных в количественной шкале, скоррелированных между собой.
- На выходе: p-мерный массив из некоррелированных главных компонент, которые представляют собой линейные комбинации исходных переменных.

Зачем, содержательно, это может понадобиться? Для решения задачи снижения размерности: вместо исходных p переменных мы можем выбрать k наиболее информативных главных компонент ($k << p$), потеряв при этом минимум информации о данных. Какие практические применения можно найти в рамках этой задачи? Вот наиболее распространённые:

- создание интегральных индексов (например, новый индекс демократии на наборе имеющихся индексов демократии, прав и свобод);
- избавление от дублирования информации (в том числе как способ борьбы с мультиколлинеарностью в множественной линейной регрессии);
- избавление от переменных с низкой вариативностью;
- избавление от шумов в данных.

Особенности метода

- На входе данные в количественной шкале (скоррелированные переменные).
- Нет особенных требований к объёму выборки.
- Не является методом моделирования, является методом перераспределения информации.
- Если сравнивать метод главных компонент с линейной регрессией, то МГК является методом построения интегральных индексов без обучения, а линейная регрессия – с обучением.

Логика метода

У нас есть массив $X = (x_1, x_2, \ldots, x_p)$, то есть массив, состоящий из вектор-столбцов $x_1, x_2, \ldots, x_p$. На его основе получаем набор из p главных компонент, таких, что произвольная j-ая главная компонента выглядит так (a_{ij} – специально подобранные коэффициенты):

$$\mathrm{PC}_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{pj}x_p.$$

Другими словами, мы должны представить исходный массив X в виде линейной комбинации исходных вектор-столбцов, таким образом, чтобы:

1. обеспечить минимальную потерю информации (разницу между старым массивом X и новым массивом X');
2. обеспечить максимальную вариативность главных компонент (первая ГК – самая информативная, вторая – менеее информативная, последняя – наименее информативная, это обеспечивается по построению).

Мы обозначили общую логику метода, далее перейдём к более глубокому изучению геометрической стороны метода главных компонент.

1.2 Метод главных компонент: какая геометрия за ним стоит

Базисные (базовые) векторы

Со школьных уроков геометрии вы, наверное, помните про базисные (базовые) единичные векторы $\vec{i}$, $\vec{j}$, через которые можно выразить любые другие векторы в декартовой системе координат.

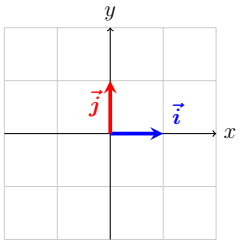


Рис. 1: Базисные (базовые) векторы $\vec{i}$, $\vec{j}$

Давайте рассмотрим пример. Вектор $\vec{a}$ имеет координаты $(2, 2)$. Его можно представить как линейную комбинацию векторов $\vec{i}$ и $\vec{j}$: сумму этих векторов, взятых с некоторыми коэффициентами. Какими именно? Давайте используем правило параллелограмма только в «обратную» сторону: зная координаты вектора, найдем, из каких векторов он получается. Спроецируем вектор $\vec{a}$ на ось x, а потом на ось y:

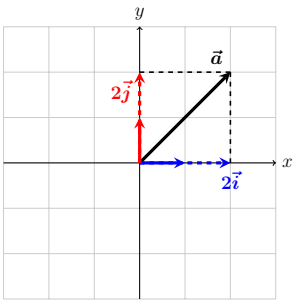


Рис. 2: Вектор $\vec{a}$

Видно, что вектор $\vec{a}$ можно получить, сложив удвоенный вектор $\vec{i}$ и удвоенный вектор $\vec{j}$. В результате получаем $\vec{a} = 2\vec{i} + 2\vec{j}$. Теперь рассмотрим вектор $\vec{b}$ с координатами $(-1, 3)$.

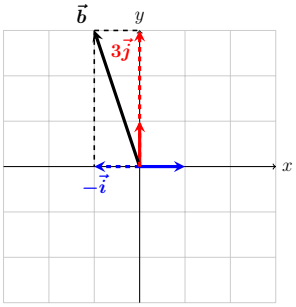


Рис. 3: Вектор $\vec{b}$

Легко заметить, что он сонаправлен с вектором $\vec{j}$, но при этом направлен в противоположную сторону по сравнению с вектором $\vec{i}$. В итоге получаем $\vec{b} = -\vec{i} + 3\vec{j}$.

А что будет, если мы выберем другие базисные векторы вместо $\vec{i}$ и $\vec{j}$? Например, возьмем вектор $\vec{i'} = \vec{i} + \vec{j}$ и вектор $\vec{j'} = \vec{j} - \vec{i}$? Мы просто повернем наши оси x и y – перейдем в новую систему координат!

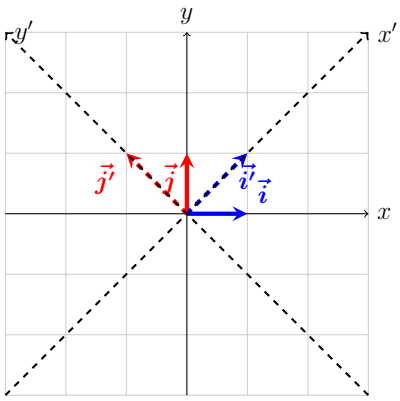


Рис. 4: Вектор $\vec{b}$

В новой системе (в новом базисе) наши векторы $\vec{a}$ и $\vec{b}$ будут иметь другие координаты. Теперь $\vec{a} = 2\vec{i'}$, а $\vec{b} = \vec{i'} + 2\vec{j'}$. Соответственно, в новой системе координат вектор $\vec{a}$ имеет координаты $(2, 0)$, а вектор $\vec{b}$ – координаты $(1, 2)$. При этом, зная, как соотносятся векторы $\vec{i'}$ и $\vec{i}$, $\vec{j'}$ и $\vec{j}$, мы всегда сможем выразить векторы $\vec{a}$ и $\vec{b}$ как в старой, так и в новой системе координат.

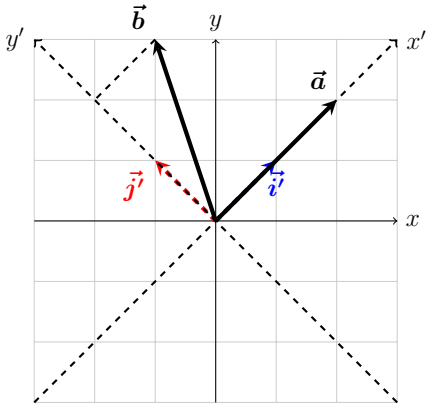


Рис. 5: Векторы $\vec{a}$ и $\vec{b}$ в новой системе координат

Идея поворота осей заложена в методе главных компонент. Мы ничего не меняем в исходных данных (только центрируем и нормируем), но зато переходим в более удобную систему координат, оси в которой расположены так, чтобы, с одной стороны, минимизировать потерю информации, а с другой стороны, обеспечить максимальную дисперсию вдоль осей. Осталось выяснить, каким образом нужно поворачивать оси, чтобы этого достичь. Другими словами, нужно найти ответы на два вопроса: каким образом описать поворот осей и какие векторы нужно взять в качестве новых базисных?

Линейные отображения и матрицы

Матрица – не просто таблица с числами. С помощью матриц задаются линейные отображения из пространства в пространство. Проще говоря, матрица (а точнее, линейное отображе-

ние, заданное матрицей) может «воздействовать» на вектор, переводя его в другой вектор в пространстве. Например, матрица (заданное ей линейное отображение) может растянуть вектор по одной оси или сразу по двум, отобразить его симметрично относительно одной оси, подвинуть в каком-то направлении на определенное число единиц, повернуть по часовой стрелке или против и так далее. Чтобы не углубляться в формальности, давайте просто рассмотрим несколько примеров.

Пусть есть вектор $v = (1, 2)$. Применим к нему линейное отображение, которое задается матрицей $A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$. Применить линейное отображение к вектору – умножить матрицу этого линейного отображения на вектор.

$$\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Что делает это линейное отображение? Растягивает вектор в два раза вдоль оси x. Давайте на это посмотрим (картина «до» и «после»):

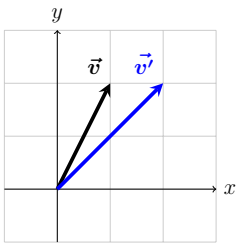


Рис. 6: Вектор $\vec{v}$ под действием линейного отображения с матрицей A

Теперь применим к вектору v линейное отображение с матрицей $B = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$. Получим следующее:

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \end{pmatrix}$$

Это линейное отображение растягивает вектор v в два раза вдоль оси x и в три раза вдоль оси y:

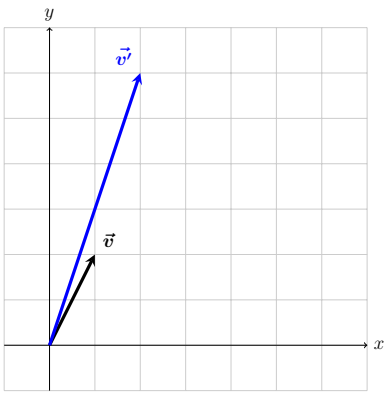


Рис. 7: Вектор $\vec{v}$ под действием линейного отображения с матрицей B

И, наконец, последний пример: линейное отображение задается матрицей $C = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Применим его к вектору v:

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

Давайте сначала посмотрим на график, и по нему поймем, что произошло.

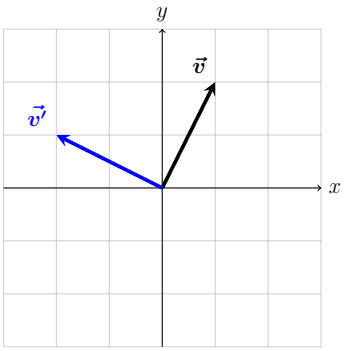


Рис. 8: Вектор $\vec{v}$ под действием линейного отображения с матрицей C

Вектор v' развернулся на 90 градусов против часовой стрелки! Конечно, это произошло не случайно. Существуют специальные матрицы – матрицы поворота. Геометрически умножение на эту матрицу означает то, что векторы поворачиваются на определенный угол по часовой стрелке или против нее. Если интересно, можете посмотреть на матрицу поворота в общем виде (здесь ϕ – это угол поворота):

$$M = \begin{pmatrix} \cos\phi & \mp\sin\phi \\ \pm\sin\phi & \cos\phi \end{pmatrix}$$

Там, где стоят знаки $\mp$ и $\pm$ подразумевается выбор знака в зависимости от направления: верхний знак, если поворот происходит против часовой стрелки, нижний – если по часовой стрелке.

Почему нам понадобилось вспоминать про матрицы? Потому, что наша задача в методе главных компонент – перейти в другую, более удобную и более экономную систему координат. А описать этот переход можно именно с помощью матрицы, которая так и назвается – *матрица перехода* (переход из одного базиса в другой). Матрицы перехода бывают разными. В случае с методом главных компонент нас будет интересовать поворот осей, значит, матрица поворота – и есть то, что нам нужно получить! На самом деле это та самая матрица, которую мы видим в R, и не случайно она называется `Rotation` (*поворот*):

```
> prcomp(dat, scale = TRUE)

Standard deviations:
[1] 1.2247449 0.7071068

Rotation:
           PC1        PC2
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
```

Итак, мы выяснили, что переход к новой системе координат (новому базису) можно задать с помощью некоторой матрицы – матрицы перехода, Осталось ответить на один вопрос: как получить эту матрицу? Сейчас выясним. Но сначала вспомним, что такое собственные значения и собственные векторы, и подумаем над геометрическим смыслом ковариации.

Собственные числа и собственные векторы

Собственный вектор матрицы M – ненулевой вектор v, который при умножении на матрицу M не меняет своего направления (остается параллельным самому себе). Такой вектор при умножении на матрицу умножается на некоторое число λ, которое называется *собственным значением* матрицы M:

$$Mv = \lambda v.$$

Давайте рассмотрим какой-нибудь пример. Возьмем матрицу $D = \begin{pmatrix} 2 & 0 \\ 1 & 4 \end{pmatrix}$. Вектор $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$ является собственным вектором этой матрицы, так как:

$$\begin{pmatrix} 2 & 0 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} -4 \\ 2 \end{pmatrix} = 2 \cdot \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

И соответствующее собственное значение $\lambda = 2$. Давайте попробуем изобразить этот вектор на графике:

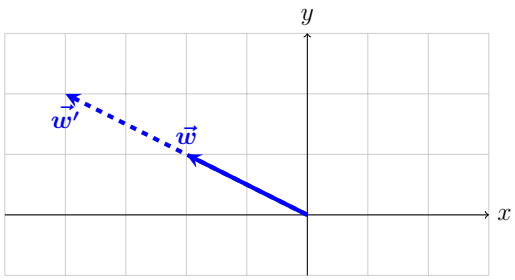
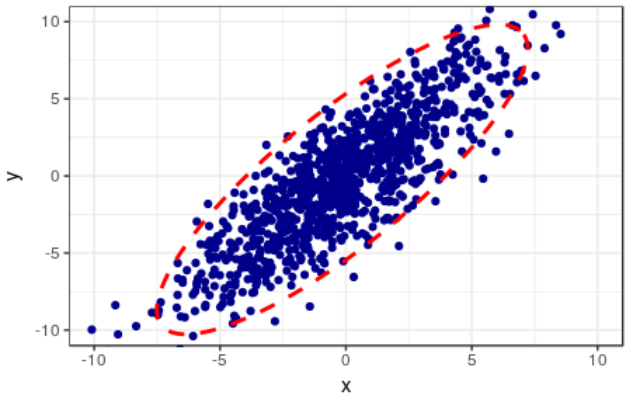


Рис. 9: Вектор $\vec{v}$ под действием линейного отображения с матрицей D

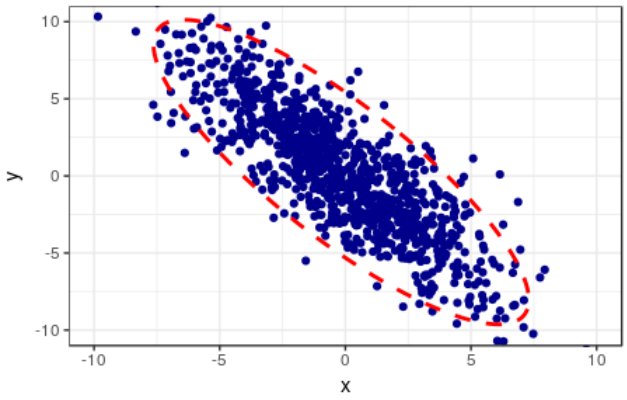
Собственное значение показывает, во сколько раз растягивается собственный вектор вдоль осей при применении к нему соответствующего линейного отображения.

Ковариационная матрица и ее геометрическая интерпретация

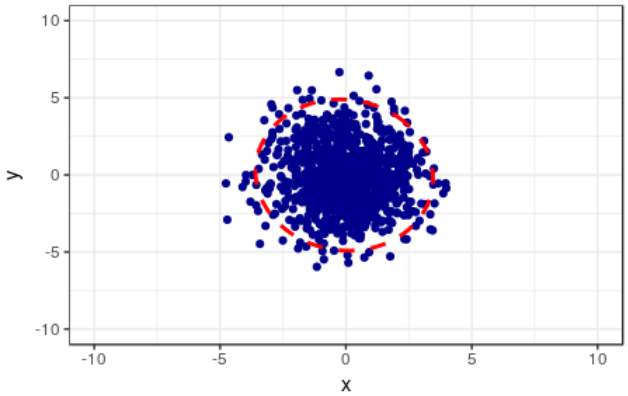
В дальнейшем нас будут интересовать не любые матрицы, а те матрицы, которые описывают наши данные. Ковариационные матрицы. Мы помним, что ковариационная матрица – это квадратная симметричная матрица, на главной диагонали которой стоят дисперсии переменных, а на побочной – значения их ковариации. Еще такая матрица должна быть неотрицательно определена – определитель такой матрицы должен быть больше или равен нулю. Давайте рассмотрим данные с разной структурой – сравним диаграммы рассеяния[1].



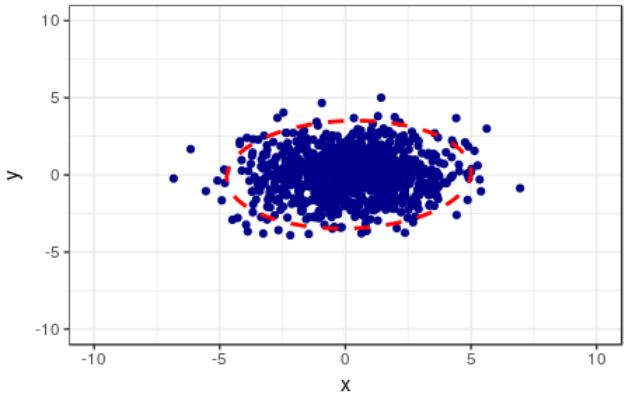
(a) Ковариационная матрица $\Sigma = \begin{pmatrix} 9 & 10 \\ 10 & 16 \end{pmatrix}$



(b) Ковариационная матрица $\Sigma = \begin{pmatrix} 9 & -10 \\ -10 & 16 \end{pmatrix}$



(c) Ковариационная матрица $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$



(d) Ковариационная матрица $\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$

Рис. 10: Диаграммы рассеяния пар показателей, описанных с помощью разных ковариационных матриц

Совместное распределение переменных можно представить в виде «облака» точек. Дисперсия отвечает за разброс значений вдоль одной из осей, а ковариация – за наклон «облака». Мы уже знаем, что матрица – это не просто таблица с числами, с помощью матрицы можно описать линейное преобразование. Какое линейное преобразование описывает ковариационная матрица? Чтобы это понять, представим себе такую ситуацию.

У нас есть нескоррелированные переменные X и Y, обе взяты из стандартного нормального распределения со средним 0 и дисперсией 1. Несложно догадаться, что таким данным соответствует ковариационная матрица такого вида:

[1]Этот раздел основан на этом материале: http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix.

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

На основе этих данных нам нужно получить данные со структурой, которая описывается другой ковариационной матрицей:

$$\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 7 \end{pmatrix}.$$

Формулируя эту же задачу в контексте диаграмм рассеяния, мы приходим к тому, что из левой картинки нужно получить правую:

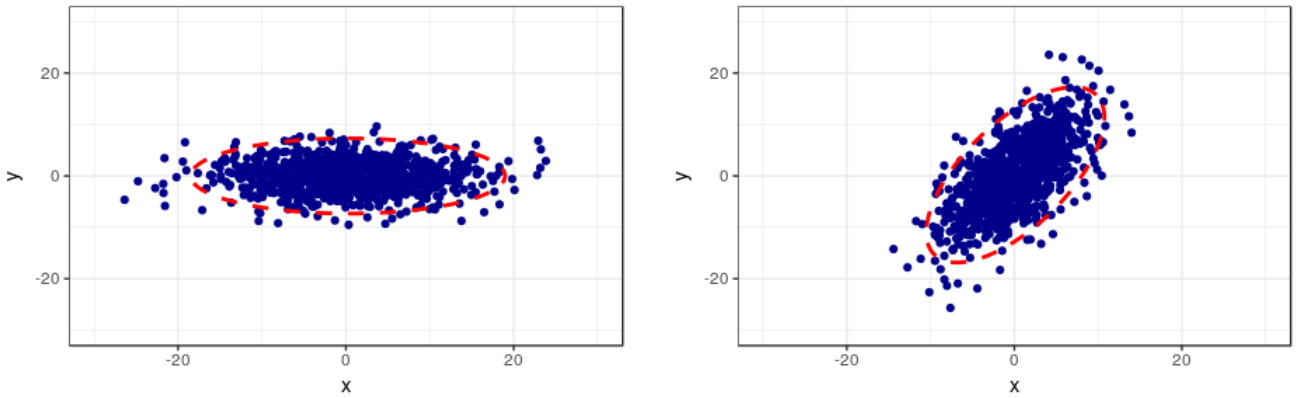


Рис. 11: «До» и «после»

Как это осуществить? По графикам видно, что «облако» точек на левой картинке нужно растянуть вдоль осей x и y, а потом его наклонить. Давайте выполним эти преобразования поэтапно. Сначала растянем «облако» точек по оси x в восемь раз, а по оси y – в три раза.

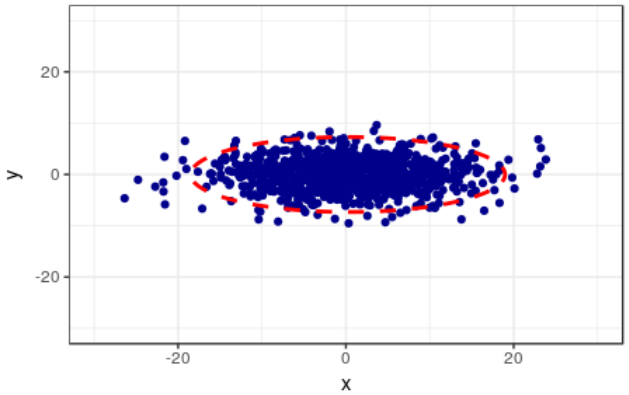


Рис. 12: Растяжение

Матрица S такого линейного преобразования (растяжения, *scale*) выглядит так:

$$S = \begin{pmatrix} 8 & 0 \\ 0 & 3 \end{pmatrix}.$$

А теперь повернем «облако» точек на нужный угол:

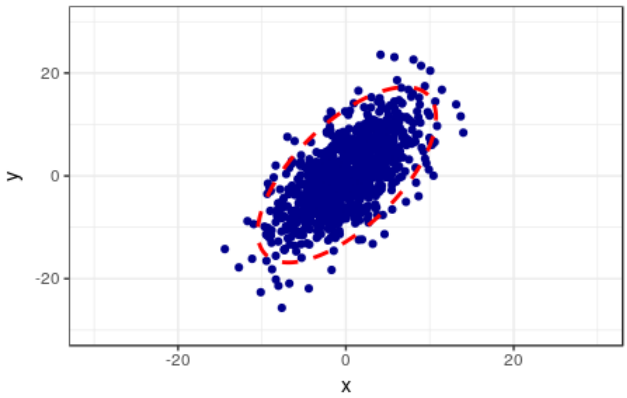


Рис. 13: Поворот

Матрица R такого линейного преобразования (поворот, *rotation*) выглядит так:

$$R = \begin{pmatrix} 0.4472136 & -0.8944272 \\ 0.8944272 & 0.4472136 \end{pmatrix}.$$

Конечно, мы выбрали эти матрицы неслучайно. Давайте сделаем следующее – перемножим матрицы:

$$\begin{pmatrix} 0.4472136 & -0.8944272 \\ 0.8944272 & 0.4472136 \end{pmatrix} \begin{pmatrix} 8 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 8 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 0.4472136 & -0.8944272 \\ 0.8944272 & 0.4472136 \end{pmatrix}^{-1} =$$
$$= \begin{pmatrix} 0.4472136 & -0.8944272 \\ 0.8944272 & 0.4472136 \end{pmatrix} \begin{pmatrix} 64 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} 0.4472136 & 0.8944272 \\ -0.8944272 & 0.4472136 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 2 & 7 \end{pmatrix}$$

Это же и есть ковариационная матрица, описывающая структуру данных, которые мы хотим получить! Действительно, ковариационную матрицу можно разложить в произведение матриц, которые будут соответствовать разным линейным преобразованиям, а именно:

$$\Sigma = RSSR^{-1}.$$

Как мы получили матрицы S и R? На самом деле, матрица S, которая «отвечает» за растяжение/сжатие вдоль осей, состоит из собственных значений ковариационной матрицы, а матрица поворота R – из ее собственных векторов единичной длины (векторы идут по столбцам матрицы). Получается, если заглянуть вглубь ковариационной матрицы конкретных переменных, мы сможем узнать, во сколько раз нужно растянуть «облако» точек, образованное значениями нескоррелированных центрированно-нормированных переменных (со средним 0 и дисперсией 1), и как его нужно повернуть, чтобы получить структуру (форму) наших данных.

Тут имеет смысл обратить внимание на следующий факт: если мы наложим на диаграмму рассеяния собственные векторы с координатами $(0.45, 0.89)$ и $(-0.89, 0.45)$[2], эти векторы будут перпендикулярны друг другу, и один из векторов будет показывать направление наибольшего разброса в данных. Теперь мы вспомнили все, что нужно, и можем перейти непосредственно к методу главных компонент.

[2]Значения из матрицы округлены.

1.3 Метод главных компонент: реализация

Утверждение. Для того, чтобы получить минимальную потерю информации об исходных данных и при этом сохранить максимальную дисперсию, нужно перейти к системе координат, в которой базисными векторами являются собственные векторы ковариационной матрицы длины 1.

Почему это так – вопрос нетривиальный, существуют разные способы доказательства этого утверждения, но, чтобы в них разобраться, нужно хорошо разбираться в линейной алгебре и математическом анализе (в частности, знать про разложение векторов по базису, умело выполнять операции с векторами и матрицами, плюс, уметь решать задачи оптимизации с ограничением). Кому интересно – можно почитать (хотя бы в Википедии) про отношение Рэлея и его связь с методом главных компонент. Еще можно пойти по другому пути и сравнивать метод главных компонент с методом наименьших квадратов: выбор положения оси, соответствующей первой главной компоненте напоминает выбор положения регрессионной прямой, так как и в том и в другом случае задачей является минимизация квадратов расстояний до прямой.

Как это утверждение связано с тем, что мы успели разобрать? Очень просто: когда мы разбирали геометрический смысл ковариационной матрицы, мы растягивали и наклоняли «облако» точек, а в методе главных компонент мы тоже будем его поворачивать, но в другую сторону. То есть выполнять обратную операцию. Давайте посмотрим на график с новыми осями:

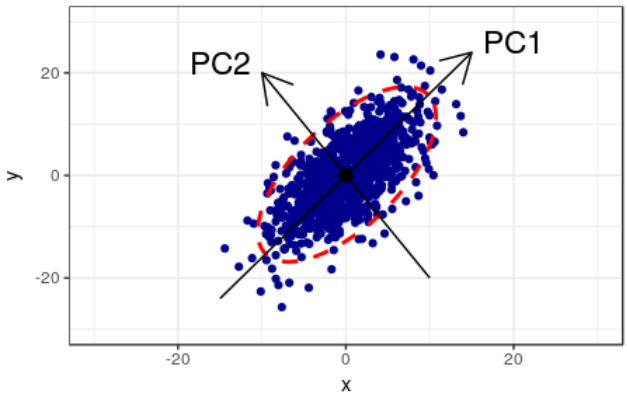


Рис. 14: Главные компоненты

Теперь развернем «облако» точек так, чтобы наши новые оси «совпали» со старыми, если перенести пересечение осей x и y в точку $(0, 0)$. То есть представим значения показателей X и Y в новой системе координат.

Что произошло? Показатели X и Y в такой системе координат никак не связаны между собой! Облако точек «лежит» горизонтально, нет никакого наклона. Как будет выглядеть ковариационная матрица X и Y? А вот так:

$$\Sigma' = \begin{pmatrix} 64 & 0 \\ 0 & 9 \end{pmatrix} = \begin{pmatrix} 8 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 8 & 0 \\ 0 & 3 \end{pmatrix}.$$

А матрица поворота?

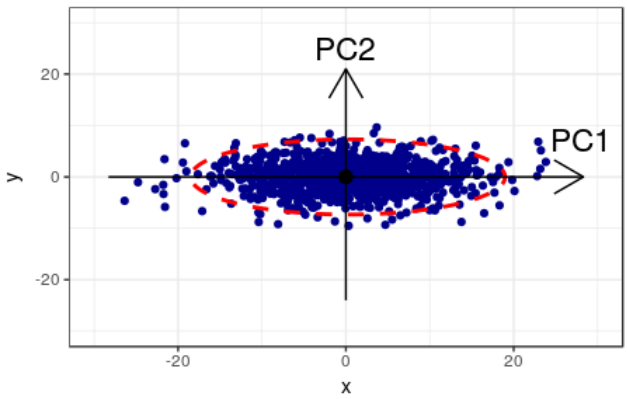


Рис. 15: Поворот

$$R = \begin{pmatrix} 0.4472136 & -0.8944272 \\ 0.8944272 & 0.4472136 \end{pmatrix}$$

Круг замкнулся (см. предыдущий раздел). Получается, что элементы на главной диагонали ковариационной матрицы соответствуют дисперсии данных вдоль осей X и Y, а собственные значения этой матрицы – дисперсии вдоль собственных векторов, то есть вдоль направлений наибольшего разброса данных, вдоль осей-главных компонент. *Главные компоненты* – новые оси, которые получились линейной комбинацией старых. Почему линейной комбинацией? Посмотрите внимательно на картинку и вспомните правило параллелограмма.

Что нам это дает? Во-первых, свойство главных компонент: ковариация между главными компонентами равна нулю. Во-вторых, гарантию, что главные компоненты не дублируют информацию – они никак не связаны. В-третьих, знания о том, какая из главных компонент важнее, то есть какая компонента сохраняет больше информации об исходных показателях. О последнем пункте поподробнее. Можно заметить, что именно вдоль более длинного вектора наблюдается больший разброс значений. Мы подошли к интересной связке: дисперсия, объясненная главной компонентой, совпадает с соответствующим ей собственным значением ковариационной матрицы.

На этом можно закончить. Осталось только описать сам алгоритм реализации метода главных компонент.

Реализация метода главных компонент

1. Шкалировать данные: вычесть из каждого значения среднее значение переменной и поделить на ее стандартное отклонение.
2. Получить ковариационную матрицу стандартизированных данных.
3. Найти собственные значения ковариационной матрицы.
4. Найти собственные векторы ковариационной матрицы, имеющие длину 1.
5. Записать матрицу из собственных векторов, найденных на предыдущем шаге. Это будет наша матрица перехода к новой системе координат, осями в которой являются главные компоненты.
6. Теперь, умножая эту матрицу на векторы, мы будем узнавать их координаты в новой системе координат.