Метод главных компонент

О чём это. Пусть у нас есть наблюдения за многомерным признаком, которые мы считаем независимыми и одинаково распределёнными:

$X_1, \dots, X_n$, где $X_i = \begin{pmatrix} X_{i1} \\ \dots \\ X_{im} \end{pmatrix}$.

Смотрите, здесь n — число наблюдений, а m — число компонент многомерного признака (число учитываемых характеристик объекта). Будем считать, что все эти характеристики количественные.

Нас расстраивает, что признак слишком многомерен. Мы хотим, чтобы его размерность была поменьше. Это не просто прихоть — многие сложности возикают именно из-за многомерности. Если бы измерений было не больше двух, значительную часть выводов можно было бы делать на глаз, посмотрев на диаграмму рассеяния или гистограмму. Во всяких регрессиях и сложных статистических процедурах потребность была бы куда меньше.

Пример задачи, в которой требуется многомерный признак свести к одномерному — рейтингование. Есть множество объектов, характеризуемых кучей показателей — например, множество регионов с различными показателями качества жизни. Чтобы упорядочить их по качеству жизни, нужно свернуть все показатели в один композитный (составной) индикатор[1] качества жизни (он же — интегральный индикатор, он же — синтетическая категория — что-то, что составляется из множества чего-то ещё). Когда такой индикатор составлен, упорядочить регионы проще простого.

Первая главная компонента

Начнём именно с такой задачи — сведение многомерного признака $X = \begin{pmatrix} X_1 \\ \dots \\ X_m \end{pmatrix}$ к одномерному индикатору. Я временно выкинул индекс i, соответствующий номеру наблюдения, потому что сейчас будет решаться чисто вероятностная задача. Будем считать, что нам не нужны наблюдения, потому что у нас есть знание настоящего распределения вектора X. А потом мы просто заменим все характеристики настоящего распределения их выборочными аналогами. В действительности, нам потребуется только ковариационная матрица.

Будем считать, что мы хотим задать величину $I = \alpha_1 X_1 + \dots + \alpha_m X_m = \alpha' X$, где $\alpha_1, \dots \alpha_m$ — веса, с которыми каждый элемент вектора X будет входить в построенный индикатор, а α — вектор, состоящий из этих весов. Веса мы не знаем и хотим определить. Будем считать, что хороший индикатор — тот, у которого большая дисперсия. Значит, веса нужно определять из задачи максимизации:

$$D(I) = D(\alpha' X) \to \max_{\alpha}. \tag{1}$$

Зачем нам нужна большая дисперсия? Для начала давайте сойдёмся на том, что композитный индикатор с нулевой дисперсией нам совершенно бесполезен — мы не сможем упорядочить регионы по величине, которая для всех регионов одинакова. Что ещё хорошего в большой дисперсии, узнаем чуть позже. Для начала попробуем решить задачу (1).

И тут мы понимаем, что задача (1) решения не имеет. Ведь дисперсию линейной комбинации $\alpha_1 X_1 + \dots + \alpha_m X_m$ можно увеличивать сколько угодно, увеличивая вес любого элемента X_i — лишь бы дисперсия этой элемента не была равна нулю. Значит, на веса нужно наложить ограничение. В методе главных компонент используется ограничение $\alpha_1^2 + \dots + \alpha_m^2 = 1$, или, в векторной форме, α'α = 1. Так мы получаем задачу условной максимизации:

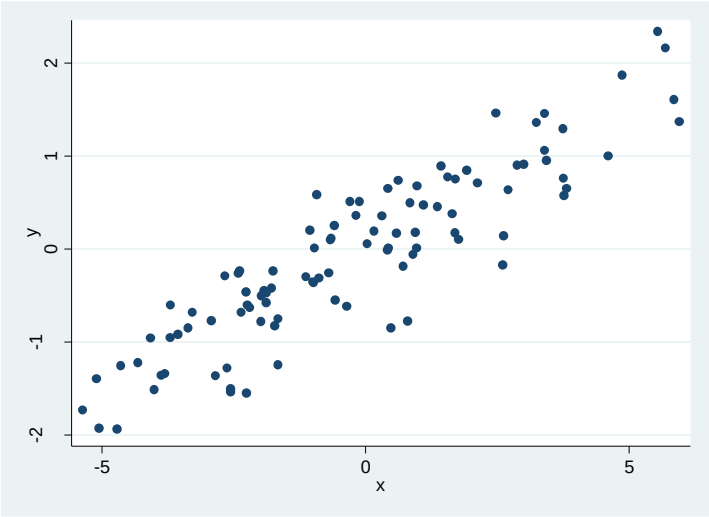
$$\begin{cases} D(I) = D(\alpha' X) \to \max_{\alpha}. \\ \alpha'\alpha = 1. \end{cases} \tag{2}$$

Решив эту задачу, мы получим композитный индикатор I, который называется *первой главной компонентой* вектора X. Мы ещё разберёмся с тем, почему она первая и каковы другие главные компоненты, но сначала я попробую нагляднее представить, что такое I.

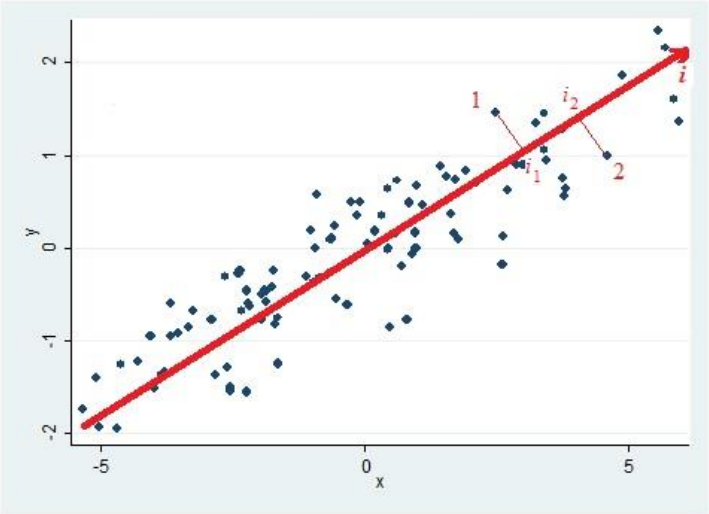
---

[1] Вместо слов «композитный индикатор» или «композитный индекс» используются и выражения «интегральный индикатор» и «синтетическая категория». В любом случае, речь идёт об индикаторе/индексе чего-то, который составляется из других индикаторов.

Представим себе набор наблюдений за двумерным признаком (X, Y) с такой диаграммой рассеяния:



Добавим к этому графику ещё одну линию — новую ось, вдоль которой разброс точек наибольший:



На этой картинке i_1 и i_2 — значения первой главной компоненты для объектов 1 и 2, иначе говоря — координаты этих объектов по оси наибольшего разброса.
Подытоживаем. Объекты в наибольшей степени отличаются друг от друга именно вдоль оси i. Можно сказать, что координата по этой оси содержит наибольшую информацию о всём двумерном признаке. Она информативнее любой другой линейной комбинации X и Y (т.е. координате по какой-либо другой оси). Мы добились этого именно потому, что ориентировались на наибольшую дисперсию и пытались добиться наибольшего разброса величины I.

Прочие главные компоненты

Чтобы разобраться с ними, придётся частично решить задачу (2).
Пусть Σ — ковариационная матрица случайного вектора X. Тогда $D(\alpha' X) = \alpha' \Sigma \alpha$ (вспоминаем свойство ковариационной матрицы: V(AX+b)=AV(X)A').
Решать задачу (2) будем с помощью функции Лагранжа:

$$L(\alpha, \lambda) = \alpha' \Sigma \alpha - \lambda(\alpha' \alpha - 1) \underset{\alpha,\lambda}{\rightarrow} max$$

Находим производные[2]:

$$\frac{\partial L(\alpha,\lambda)}{\partial \alpha} = 2\Sigma\alpha - 2\lambda\alpha;$$

[2] Допускаю, что вы никогда не дифференцировали функцию по векторному аргументу. Предлагаю просто поверить мне, что производные выглядят так. Кое-какие правила дифференцирования можно найти в книжке Магнуса, Катышева, Пересецкого, которую я рекомендовал для изучения регрессионного анализа, но там они даются без объяснений – не многим лучше, чем просто предложить поверить.

$$\frac{\partial L(\alpha,\lambda)}{\partial \lambda} = \alpha'\alpha - 1.$$

Приравняв первую из них к нулю, получаем: $\Sigma\alpha = \lambda\alpha$.
Тут нужно сделать экскурс в линейную алгебру и вспомнить, что такое собственные векторы и собственные числа.

Число λ называется *собственным числом*, а вектор α — соответствующим этому числу *собственным вектором* квадратной матрицы Σ, если выполняется равенство $\Sigma\alpha = \lambda\alpha$. Далее будем рассматривать только собственные векторы с единичной длиной: $\alpha'\alpha = 1$ (это типичная в алгебре нормировка, без неё будет бесконечно много собственных векторов).

У матрицы размера m x m всего m собственных векторов и m соответствующих им собственных чисел. Упорядочим собственные числа в порядке убывания: $\lambda^1 > \lambda^2 > \ldots > \lambda^m$ (сверху – индекс, а не степень). Соответствующие им векторы будем обозначать также верхним индексом: $\alpha^1, \alpha^2, \ldots, \alpha^m$ — это чтобы не путать с компонентами отдельного вектора, которые я раньше уже обозначал нижним индексом. Случайная величина $I^j = \alpha^{j\prime}X$ называется *j-й главной компонентой* вектора *X*. Чтобы понять, что это такое, вспоминаем *свойства собственных векторов и чисел*:

1º. $\alpha^{j\prime}\Sigma\alpha^j = \lambda^j$. То есть дисперсия каждой главной компоненты равна соответствующему собственному числу.
Доказательство: ◄ $\alpha^{j\prime}\Sigma\alpha^j = \alpha^{j\prime}\lambda^j\alpha^j = \lambda^j\alpha^{j\prime}\alpha^j = \lambda^j$ ►

2º. Собственные векторы α^j, α^k, соответствующие *различным* собственным числам λ^j и λ^k ортогональны: $\alpha^{j\prime}\alpha^k = 0$.
Доказательство писать лень уже, хотя оно короткое. Про то, что происходит с собственными векторами, соответствующими одинаковым собственным числам, тоже писать не хочу — давайте считать, что все собственные числа различны[3].

Остальные свойства собственных векторов и чисел не будем впоминать, потому что сейчас они не пригодятся.

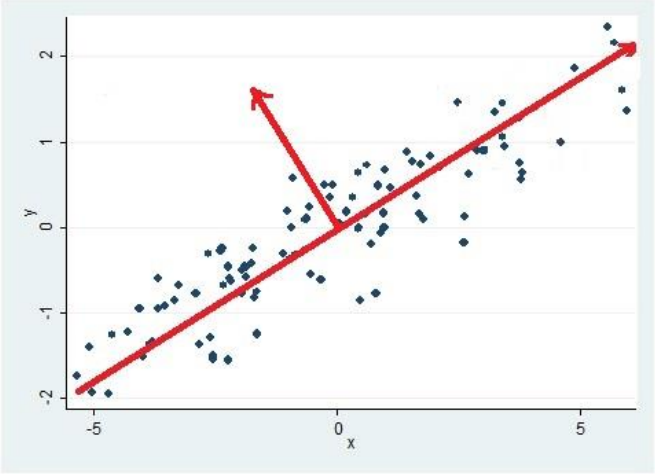
Мораль из указанных двух свойств:
Все главные компоненты ортогональны. Можно сказать, что это проекции вектора X на ортогональные подпространства и, как это всегда случается с проекциями на ортогональные подпространства, они не коррелированы.
Мы упорядочили сбственные числа по убыванию, а они равны дисперсиям главных компонент. Отсюда смысл:
Первая главная компонента — это линейная комбинация составляющих вектора X с наибольшей дисперсией (об этом уже говорилось).
Вторая главная компонента — линейная комбинация составляющих вектора X с наибольшей дисперсией среди всех линейных комбинаций, ортогональных первой главной компоненте.
Третья главная компонента — линейная комбинация составляющих вектора X с наибольшей дисперсией среди всех линейных комбинаций, ортогональных первым двум главныи компонентам.
И так далее.
Последняя главная компонента — линейная комбинация составляющих вектора X с наименьшей дисперсией.

При этом мы ограничиваемся только линейными комбинациями с ограничениями на веса: $\alpha'\alpha = 1$.

Может возникнуть вопрос: как всё это выросло из задачи максимизации дисперсии (2)? На самом деле, из неё вытекает только первая главная компонента. Остальные возникают лишь из условия первого порядка (равенства первых производных нулю), которое выполняется и для максимума, и для минимума, поэтому комбинация с минимальной дисперсией тоже входит главные компоненты.

Главные компоненты на картинке выглядят так:

[3] Тем, кто недоволен, короткая напоминалка: собственные векторы, соответствующие одинаковым собственным числам, тоже можно сделать ортогональными (процедура Грамма–Шмидта).



К предыдущему рисунку здесь добавляется ещё одна красная ось, ортогональная оси наибольшего разброса. Координата каждой точки по этой второй оси равна значению второй главной компоненты. По добавленной оси разброс точек, наоборот, минимален.

Если представить себе ряд наблюдений за трёхмерным признаком с графиком рассеяния в виде батона, то ось первой главной компоненты будет направлена вдоль батона, вторая — вширь, третья — ввысь.

Замечание. Дисперсия изменится, если поменять единицы измерения признака. Чтобы главные компоненты не были чувствительны к выбору единиц измерения, все компоненты многомерного признака можно поделить на их стандартные отклонения. Это уничтожает размерность и эквивалентно расчёту весов по собственным векторам корреляционной матрицы вместо ковариационной (ведь корреляция — это ковариация величин, поделённых на их стандартные отклонения). По умолчанию статистические пакеты обычно рассчитывают главные компоненты именно по корелляционной матрице, так что результаты не зависят от размерности.

Вклад каждой компоненты в общий разброс и выбор нужного числа главных компонент

Как показатель информативности каждой главной компоненты используют отношение дисперсии этой компоненты к сумме дисперсий всех компонент: $\frac{\lambda^j}{\lambda^1+\cdots+\lambda^m}$. Когда решается задача снижения размерности, то из всех главных компонент оставляются только наиболее информативные. Например, можно поставить цель — ограничиться теми главными компонентами, на долю которых приходится 90% всего разброса. Но вообще-то людям не свойственно иметь в голове такие чёткие критерии, поэтому на практике пользуются другими правилами.

Бессмысленная байка. Однажды в метро я задумался: как так выходит, что приезжает поезд, из вагона выходит целая толпа людей, а он остаётся забитым? Ведь те люди, которые выходили из вагона, не могли там поместиться. Тогда я как раз изучал многомерный статистический анализ и решил, что в метро настоящих людей незаметно подменяют их главными компонентами, которые гораздо компактнее и почти идентичны исходным людям (например, содержат 99.9% информации о них). Однако часть информации всё же теряется, поэтому в метро искажаются – это объясняло, почему один стоит на платформе и разговаривает с рельсами, а другой в вагоне пытается что-то втолковать своему отражению в стекле. И прочее в том же духе, что в метро наблюдается нередко. У этих людей часть была потеряна.
Простите за злоупотребление вашим вниманием, я возвращаюсь к делу.

Одно из распространённых правил отбора компонент — *критерий Кайзера*. Согласно этому критерию, нужно оставлять те компоненты, которые имеют дисперсию больше единицы (т.е. дисперсия которых больше дисперсии отдельной компоненты вектора X — мы ведь предварительно делим все компоненты на их стандартные отклонения, так что дисперсия каждой равна единице). Чуть позже мы рассмотрим ещё *критерий каменистой осыпи*.

Пример использования метода главных компонент

Этот пример взят из книжки Нэреша Малхотры «Маркетинговые исследования. Практическое руководство» и опирается на искусственные данные, но я встречал почти идентичные реальные исследования — просто реальных данных у меня на руках нет, а искусственные есть.

В файле «Зубная паста (Малхотра).ods» приведены результаты опроса 30 потребителей зубной пасты. Их попросили указать степень согласия со следующими утверждениями по семибалльной шкале (1 — абсолютно не согласен, 7 — полностью согласен):

V1	важно приобрести зубную пасту, которая предотвращает развитие кариеса;
V2	мне нравится зубная паста, которая придаёт зубам белизну;
V3	зубная паста должна укреплять дёсны;
V4	я предпочитаю зубную пасту, которая освежает дыхание;
V5	предотвращение порчи зубов не является важным преимуществом зубной пасты;
V6	самой важной причиной покупки данной зубной пасты служит её способность улучшать внешний вид зубов.

Рассчитаем корреляционную матрицу для этих величин.

```
. correlate
(obs=30)

             |       v1       v2       v3       v4       v5       v6
-------------+------------------------------------------------------
          v1 |   1.0000
          v2 |  -0.0405   1.0000
          v3 |   0.8731  -0.1428   1.0000
          v4 |  -0.0862   0.5905  -0.2478   1.0000
          v5 |  -0.8576  -0.0066  -0.7778  -0.0066   1.0000
          v6 |   0.0042   0.7126  -0.0181   0.6405  -0.1364   1.0000
```

Как видно, есть две группы тесно коррелирующих величин. В первую входят V1, V3 и V5, во вторую — V2, V4 и V6. Теперь рассчитаем главные компоненты и их вклад в общую дисперсию.

```
. pca v1-v6

Principal components/correlation                  Number of obs    =         30
                                                  Number of comp.  =          6
                                                  Trace            =          6
    Rotation: (unrotated = principal)             Rho              =     1.0000

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      2.72543      .441115             0.4542       0.4542
           Comp2 |      2.28431      1.84984             0.3807       0.8350
           Comp3 |      .434467      .145232             0.0724       0.9074
           Comp4 |      .289235       .10584             0.0482       0.9556
           Comp5 |      .183395      .100229             0.0306       0.9861
           Comp6 |     .0831663            .             0.0139       1.0000
    --------------------------------------------------------------------------

Principal components (eigenvectors)

    ------------------------------------------------------------------------------------------
        Variable |    Comp1     Comp2     Comp3     Comp4     Comp5     Comp6 | Unexplained
    -------------+--------------------------------------------------------------+-------------
              v1 |   0.5615    0.1712   -0.0770    0.2599    0.1758    0.7423 |           0
              v2 |  -0.1776    0.5467    0.5354    0.6019   -0.0465   -0.1359 |           0
              v3 |   0.5665    0.0919    0.1766   -0.1453    0.5896   -0.5202 |           0
              v4 |  -0.2108    0.5149   -0.7648    0.1369    0.2553   -0.1470 |           0
              v5 |  -0.5240   -0.2376    0.1817    0.0217    0.7440    0.2863 |           0
              v6 |  -0.1148    0.5846    0.2415   -0.7279   -0.0232    0.2373 |           0
    ------------------------------------------------------------------------------------------
```

В верхней таблице приведены характеристики разброса каждой главной компоненты: Eigenvalue (собственное число) — дисперсия главной компоненты, Difference — разность дисперсий (показывает, насколько эта главная компонента «информативнее» следующей), Proportion — доля дисперсии компоненты в общей дисперсии, Cumulative — накопленная доля. Например, на первые две главные компоненты приходится 83.5% суммарной дисперсии. Если пользоваться критерием Кайзера, то оставлять нужно первые две компоненты.

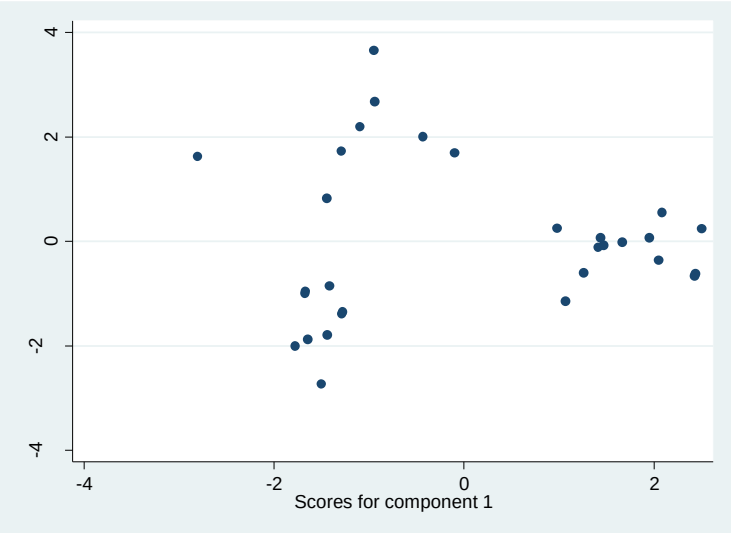
Что гораздо важнее критерия Кайзера — именно первые две главные компоненты имеют осмысленную интерпретацию. Нижняя таблица содержит веса каждой величины из V1–V6 в каждой из компонент. Получается, первые две компоненты выглядят так:

$$I^1 = 0.56V_1 - 0.18V_2 + 0.57V_3 - 0.21V_4 - 0.52V_5 - 0.11V_6;$$
$$I^2 = 0.17V_1 + 0.55V_2 + 0.09V_3 + 0.51V_4 - 0.24V_5 + 0.58V_6.$$

В первую главную компоненту с относительно большими весами вошли величины V1, V3, V5, во вторую — V2, V4, V6. Можно сказать, что первая компонента отражает важность, которую для респондента имеет способность зубной пасты улучшать здоровье, а вторая главная компонента отражает важность, которую для респондента имеет способность зубной пасты улучшать какие-то внешние проявления. Можно условно назвать их так: I^1 — важность здоровья, I^2 — важность красоты.

Зачем нам это? Например, вот зачем. Посмотрим теперь на диаграмму рассеяния в осях первых двух компонент (по горизонтали будет важность здоровья, по вертикали — красоты).

```
. predict health beauty
. scatter beauty health
```



На графике можно выделить три группы респондентов. Справа находятся те, кто придаёт значение способности зубной пасты улучшать здоровье. Сверху — те, кто придаёт значение способности улучшать внешние проявления. Внизу — те, кто просто покупает зубную пасту и ни о чём особо не беспокоится. Есть ещё один отщепенец слева, презирающий вклад зубной пасты в здоровье человека — вряд ли стоит выделять его в отдельный кластер.

Мораль: благодаря снижению размерности с шести до двух измерений мы получили возможность наглядно представить предпочтения потребителей зубной пасты и классифицировать их.

Пожалуй, достаточно. Я, правда, так и не написал здесь, как найти собственные числа и векторы, но про это можно почитать в книжках по линейной алгебре.

В основном, я ориентировался на книжку G.S. Maddala, “Introduction to Econometrics”[4] и на то, что мне рассказывали С.А. Айвазян и Т.А. Дуброва — мои преподаватели, которым спасибо.
Всё.

P.S. Возможно, для ФКН лучше подошёл бы пример использования меотда главных компонент для обработки изображений. Он действительно там используется, но у меня пока нет хорошего подробного примера. Если у вас такой пример есть и вы не против поделиться, присылайте.

[4] В отличие от регрессионного анализа, для изучения которого действительно хорошо читать книжки по эконометрике, МГК можно — а может и нужно — изучать и по другим источникам. Я использовал книгу Маддалы, потому что привык к ней (я сам учился на кафедре математической экономики и эконометрики).