# (Изложение материала в большей степени основано на: *Айвазян С.А, Мхитарян В.С.* Прикладная статистика и основы эконометрики, 1998 г. Глава 13.2: Метод главных компонент)

# Снижение размерности многомерного признака

# Пусть матрица *Х*(*n*×*p*) содержит результаты наблюдений за *n* объектами. Каждое наблюдение – многомерный показатель ***х*** ∈, регистрируемый на каждом из *n* объектов. Матрицу *Х* называют матрицей «объект-свойство», каждая строка матрицы содержит значения *р* признаков, характеризующие объект (или определяющие *свойство объекта*).

Естественно (почему естественно – вопрос к аудитории) стремление представить каждое из наблюдений, ***х***, в виде вектора **z** c существенно меньшим, чем *р*, числом признаков, или *компонент*, *p*′, т.е. в виде вектора **z**∈: *p*′<<*p*.

Новые признаки, *z*1,…,*zp*′, могут *выбираться* из числа исходных признаков (например, на основе регрессионной модели), или *определяться* по какому-либо правилу по совокупности исходных признаков. Новые признаки должны быть в определенном смысле наиболее информативными. Очевидно, что при решении (общей) задачи классификации наибольший интерес представляют признаки, демонстрирующие наибольшие изменения своих значений при переходе от одного объекта к другому. Это необходимо учитывать при формировании критерия информативности.

Предпосылки для перехода к меньшему числу признаков:

– дублирование информации, доставляемой сильно взаимосвязанными признаками;

– неинформативность некоторых признаков;

– возможность агрегирования по некоторым признакам.

## Поставим задачу: по результатам наблюдений, представленных матрицей «объект-свойство», найти преобразование признаков, (***x***)=, доставляющее максимум некоторому критерию информативности :

|  |  |
| --- | --- |
|  | (1) |

# Метод главных компонент (Principal Component Analysis, РСА) является одним из активно применяемых на практике методов *определения* (т.е. *построения*) новых признаков.

При *построении* новых признаков в РСА в качестве критерия информативности рассматривается доля наблюдаемой дисперсии исходных признаков (их количество равно *р*), воспроизводимая новыми признаками (их количество равно

**Основные понятия, определения и замечания**

В PCA семейство преобразований исходных признаков, , сужается до семейства нормированных ортогональных линейных операторов размерности :

|  |  |
| --- | --- |
|  | (3) |

где *I* (*p*×*p*) – единичная матрица, , – строки матрицы *L*(*p*×*p*).

Тогда вектор *искомых* переменных определяется как

**z** = *L* ***x***, (4)

а каждая компонента вектора записывается в виде линейной комбинация:

, (5)

(*Заметим*, что, вообще-то, размерность построенного по правилу (4) вектора **z** будет совпадать с размерностью исходного вектора ***х***!)**.**

В РСА линейные комбинации (5) строятся таким образом, чтобы *первые* из них воспроизводили *максимальную долю дисперсии* наблюдений.Тогда число новых признаков,, будет определяться на основании некоторого правила, опирающегося на критерий информативности (2) (см. пункт **Выбор** .

Суть РСА состоит в поиске таких линейных комбинаций исходных признаков (5), что:

|  |  |
| --- | --- |
|  | (6) |
|  | (7) |
|  | (8) |

Отметим, что в силу ортогональности строк матрицы *L* условие (7) заведомо выполняется.

Таким образом, 1-ой главной компонентой, **,** исследуемой системы показателей называется такая (нормированно-центрированная) линейная комбинация этих показателей, которая среди всех прочих (нормированно-центрированных) линейных комбинаций переменных обладает *наибольшей дисперсией*. Тогда *k*-я главная компонента – некоррелированная с (*k*–1) предыдущими главными компонентами и среди всех прочих (нормированно-центрированных) линейных комбинаций переменных обладает наибольшей дисперсией.

***Замечание!!!*** *Обязательным* условием применения РСА является *центрированность* данных. Обычно данные не только центрируют, но и нормируют. Если параметры распределения вектора ***х*** неизвестны, то по исходным данным (матрице *Х*) находят выборочные оценки вектора средних и матрицы ковариаций и переходят к нормированно-центрированным (*стандартизованным*) данным:

|  |  |
| --- | --- |
|  |  |

(или только центрированным: ).

**Вычисление главных компонент**

Согласно соотношению (5), **.**

Так как данные центрированы, т.е. имеем: . Вычислим дисперсию 1-й главной компоненты:

В соответствии с условием (6) и учитывая, что , для вычисления 1-й главной компоненты требуется решить оптимизационную задачу вида:

.(9)

Для решения задачи (9) воспользуемся методом Лагранжа (строим лагранжиан и приравниваем гдадиент по к нулю):

Получаем:

|  |  |
| --- | --- |
|  | (10) |

Т.к. по условию (см. (3))

(11)

Таким образом задача (9) сводится к проблеме нахождения собственных чисел и (нормированных) собственных векторов матрицы . Поскольку матрица – симметричная, неотрицательно определенная, уравнение (11) имеет *р* вещественных неотрицательных корней, т.е собственных чисел матрицы

*.*

Учитывая, что **,** , а из равенства (10) ⇒ , получаем:

Таким образом, для обеспечения максимальной величины дисперсии 1-й главной компоненты, , нужно выбрать из *р* собственных значений матрицы наибольшее, т.е. , где – нормированный собственный вектор, соответствующий и ***найти***

Приходим к общему правилу определения *k*-й главной компоненты и её дисперсии:

, (12)

, (12′)

где – нормированный собственный вектор, соответствующий , *k.*

**Основные числовые характеристики и свойства главных компонент**

1. (здесь – матрица ковариаций вектора **z**)  
   ⇒
2. суммы дисперсий иравны (т.е. условие (8) выполнено):
3. обобщенная дисперсия исходных признаков равна обобщенной дисперсии главных компонент :

**Выбор**

Исходя из соотношения (12′), критерий информативности, можно представить как

. (13)

Если данные нормированы, то есть , выражение (13) принимает вид:

. (14)

Опираясь на выражения (13), (14), можно разумно определить число компонент, которое целесообразно оставить в рассмотрении, сократив при этом размерность исследуемого признака. На практике применяют, например, одно из правил:

1. оставляют факторы с собственными числами, большими 1;

или:

1. для каждого фактора оценивают долю дисперсии, воспроизводимую этим фактором. Правило определяется уровнем (порогом) для минимальной доли воспроизводимой дисперсии (например, 5%, 10%). Правило 1) эквивалентно данному правилу для уровня (100/*р*)%, *р* – число исходных признаков.

**Матрица нагрузок** – это матрица *А* = (*aij*), «нагрузок» главных компонент на исходные признаки:

.

(Полагаем, что данные **стандартизованы**: *M*)

Определим вектор *нормированных* главных компонент: **.**

Тогда .

Учитывая, что (см. (3)), из соотношения **z** = *L****x*** получаем:

.

Тогда

, (15)

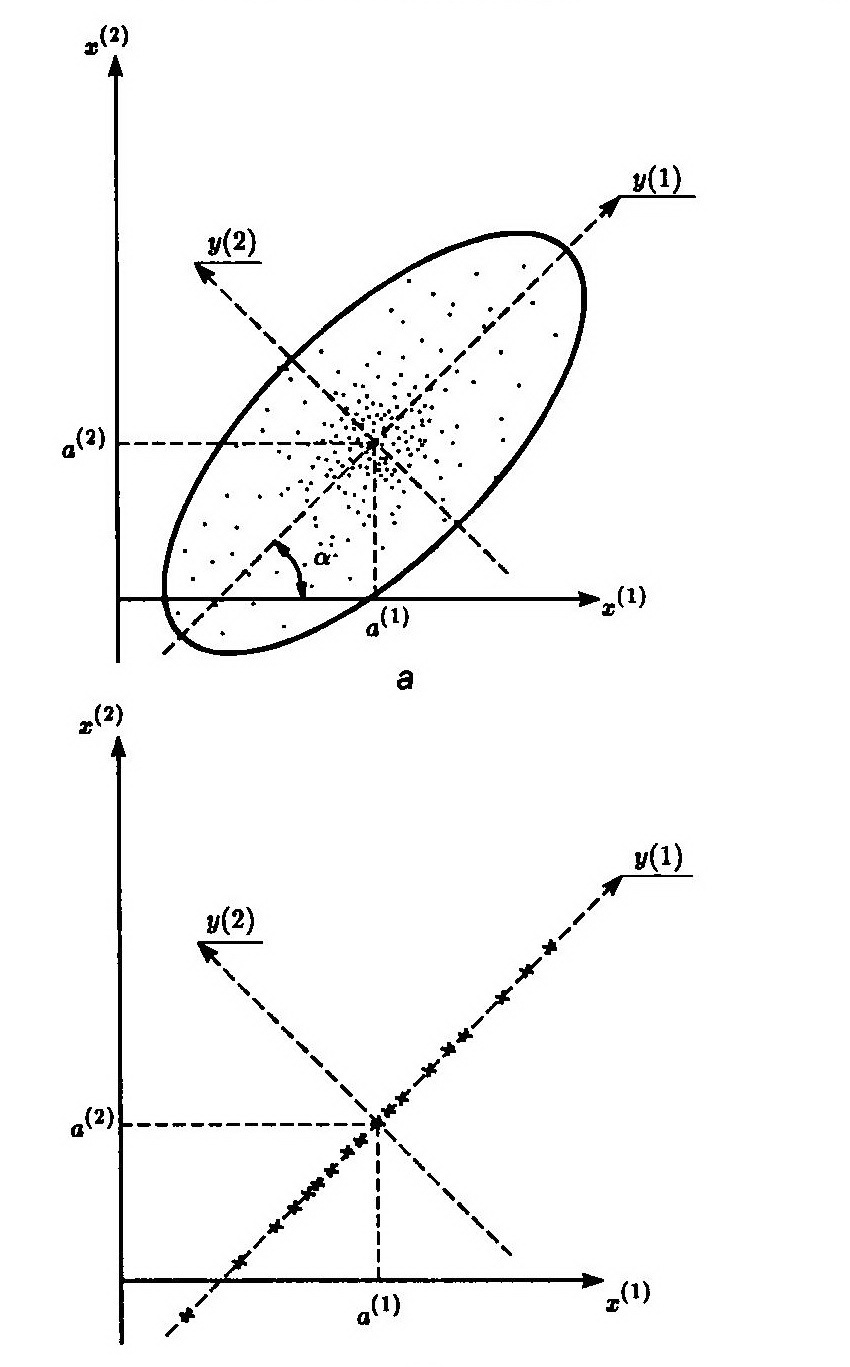
откуда следует, что:

1. определяет удельный вес влияния *j-*ой нормированной главной компоненты на *i*-й признак;
2. *,* если переменная ***х*** стандартизована (доказать самост.)
3. ,

Эти свойства используют при содержательной интерпретации главных компонент. Например, соотношение (15) дает основание придавать главной компоненте содержательный смысл, соответствующий признаку , для которого достигает максимального значения (при условии, что ||>0.6).

**К геометрической интерпретации РСА** (см., например, Айвазян С.А, Мхитарян В.С. Прикладная статистика и основы эконометрики, гл.13.2 Метод главных компонент)

РСА – это метод представления данной последовательности наблюдений в *новой системе координат*. Геометрическая интерпретация РСА иллюстрируется обычно на примере двумерного нормального распределения для коррелированных с.в. Х1 и X2. В этом случае линии уровня плотности вероятности имеют форму эллипсов с осями р1 и р2. Дисперсия в направлении первой, *главной* оси р1, максимальна. Например, для трехмерного распределения главная ось – наибольший диаметр эллипсоида рассеяния исследуемых наблюдений, вторая ось пройдет по наибольшему диаметру эллипсоида в плоскости, перпендикулярной первой оси, третья ось будет самой короткой. Понятие главных компонент относится не только к нормальным распределениям! В общем случае главная ось задается линией, вдоль которой сумма квадратов расстояний до всевозможных точек минимальна (см. рис.). Для сравнения: в МНК минимизируется сумма квадратов расстояний по линии, *параллельной* оси у.



На последнем рисунке главные компоненты – y(1) и y(2) ( вместо )

Что такое