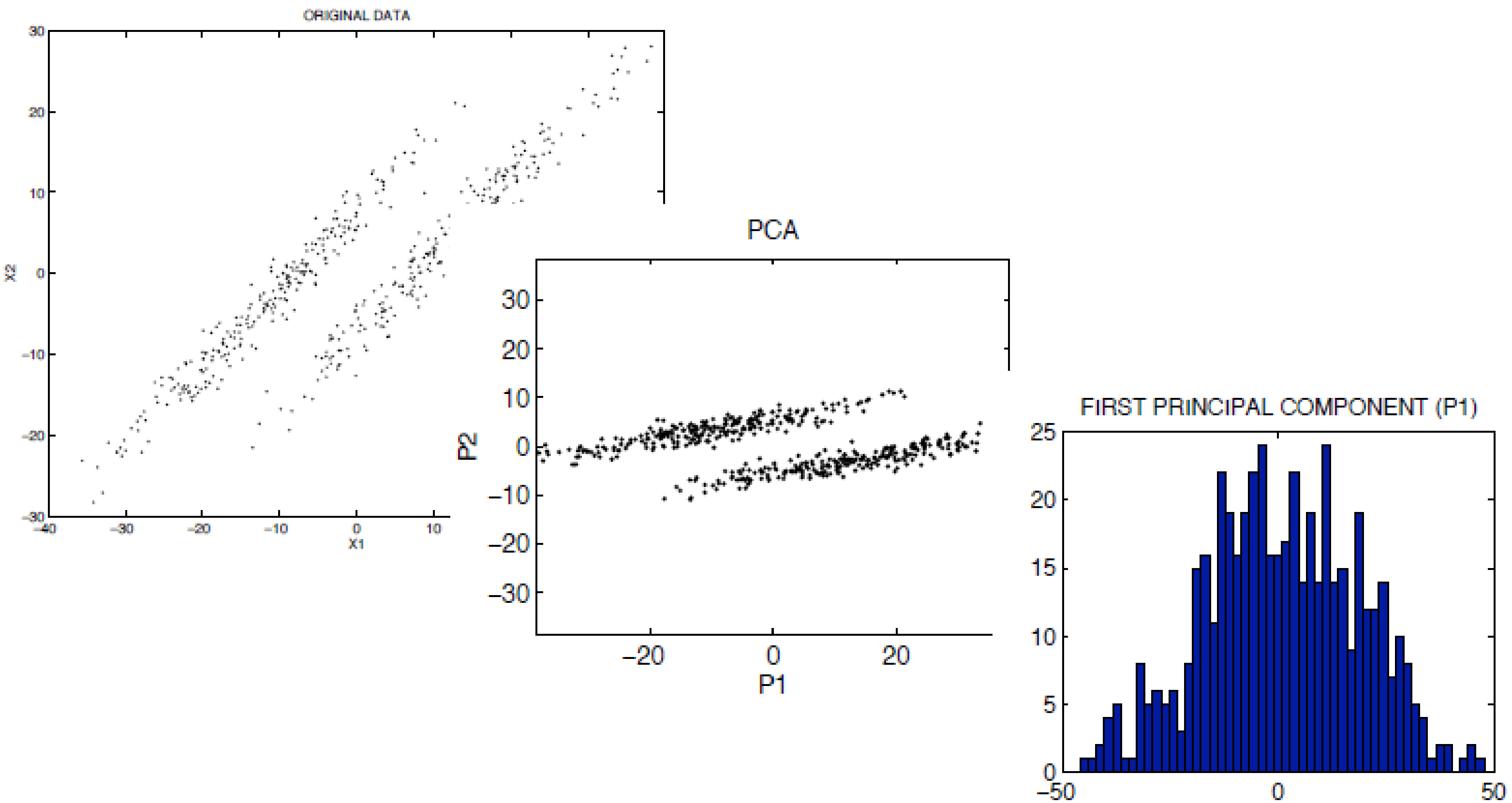
Условия применимости PCA

Общие условия

- Наличие выбросов
  - Исключить выбросы
- Наличие нулей (часто это – пропущенные значения)
  - Если много, выполнить трансформацию данных
  - Если очень много, удалить такие переменные из анализа
- PCA – линейный метод
  - Наличие линейных связей между переменными

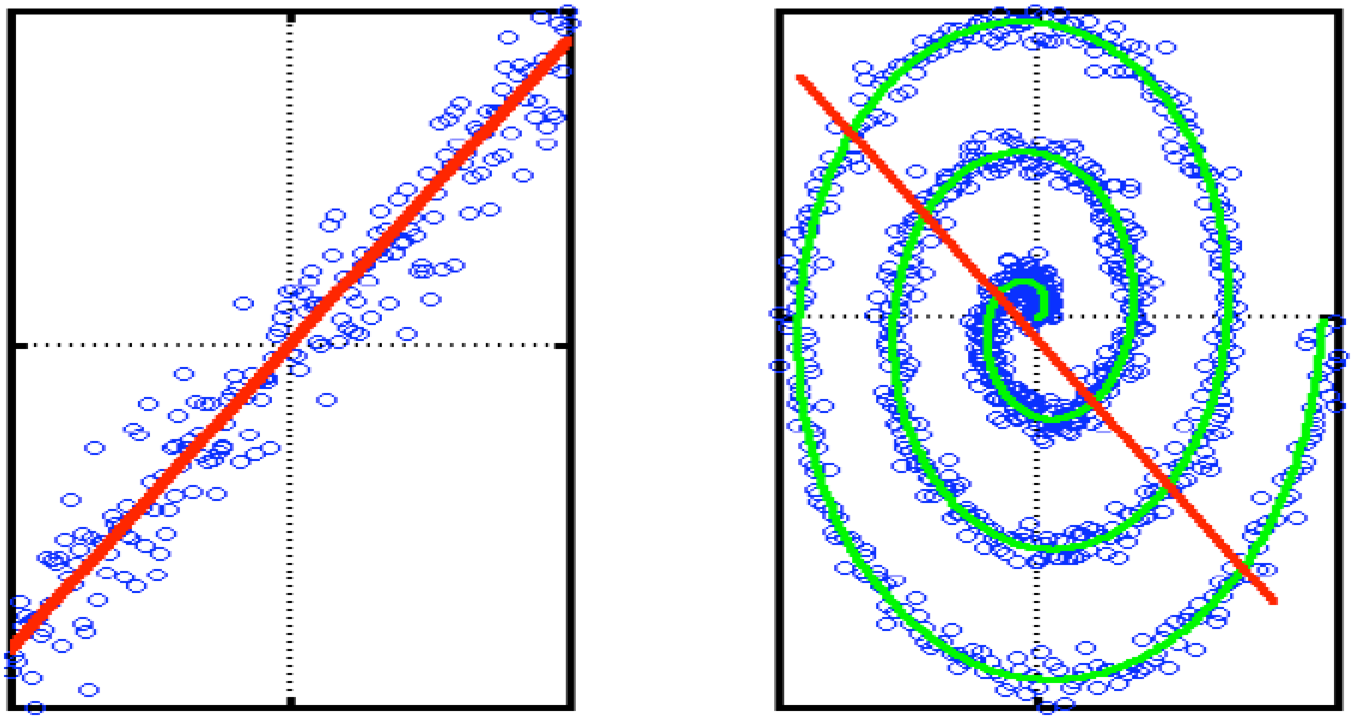
Two clusters

- The PCA fails to separate the clusters (you don't see cluster structure from the 1D visualization, lower right)



Источник: Jaakko Peltonen, Spring 2014

Nonlinear data

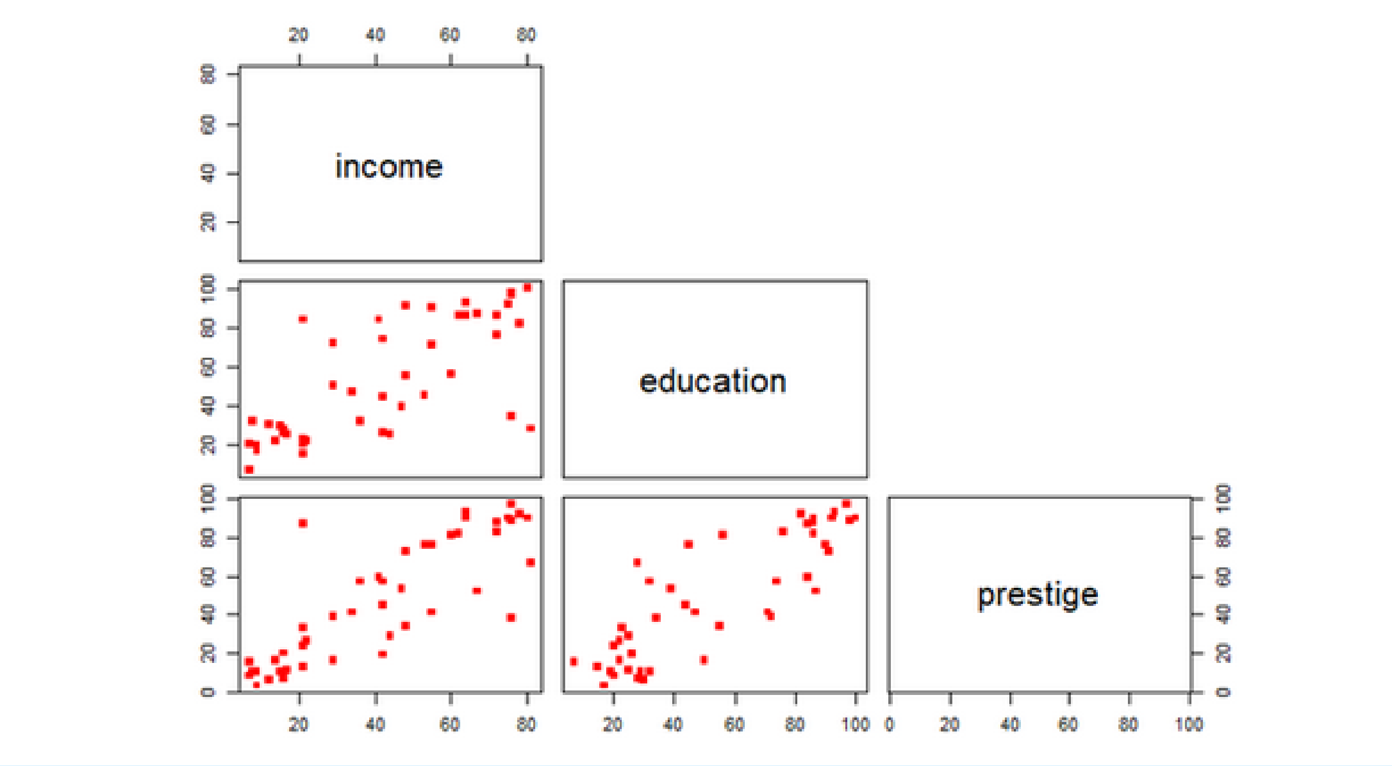


The first principal component is given by the red line. The green line on the right gives the "correct" non-linear dimension (which PCA is of course unable to find).

Источник: Jaakko Peltonen, Spring 2014

Применять PCA? Графически

- Визуализировать корреляционную матрицу.
- Диаграмма рассеяния.



Условия применимости PCA

- Тест сферичности Бартлетта
- Индекс KMO

  - Встроены в статистические пакеты (SPSS)
  - Реализованы в R и Python

Тест сферичности Бартлетта

- Основан на **корреляционной** матрице.
- Тест Бартлетта сравнивает наблюдаемую корреляционную матрицу с единичной матрицей.
- Если переменные полностью ***коррелированы***, достаточно только одного фактора.
- Если переменные ***ортогональны***, нам нужно столько же факторов, сколько и переменных. В этом случае корреляционная матрица - единичная. PCA бесполезен.
- Если большинство элементов корреляционной матрицы близки к нулю, то есть корреляционная матрица близка к единичной, PCA бесполезен.
- Нужно подтвердить, что наши гипотезы статистически значимы.
- **Применить тест сферичности Бартлетта**.

	Population	School	Employment	Services	HouseValue
Population	1.00000000	0.00975059	0.9724483	0.4388708	0.02241157
School	0.00975059	1.00000000	0.1542838	0.6914082	0.86307009
Employment	0.97244826	0.15428378	1.0000000	0.5147184	0.12192599
Services	0.43887083	0.69140824	0.5147184	1.0000000	0.77765425
HouseValue	0.02241157	0.86307009	0.1219260	0.7776543	1.0000000

Корреляционная матрица

(социально-экономические признаки)

- Некоторые переменные взаимосвязаны
  - Численность населения и занятость: 0,97;
  - стоимость школьного обучения и стоимость дома: 0,86.
- Возможно, стоит применить PCA.
- Подтвердим это с помощью **статистической** гипотезы – тест Бартлетта.

Тест сферичности Бартлетта

- Проверяет, значительно ли отличается наблюдаемая корреляционная матрица $\mathbf{R}_{pxp}$ от единичной матрицы. Здесь **p** – количество переменных (признаков).
- Нулевая гипотеза H0: **переменные ортогональны**.
- Если нулевая гипотеза отклоняется, можно применять PCA для сжатия пространства исходных переменных.
- Чтобы измерить общую связь между переменными, вычисляется определитель корреляционной матрицы | R |.

  При H0, $|R| = 1$, если же переменные сильно коррелированы, $|R| \approx 0$.
- Статистика теста Бартлетта показывает, в какой степени мы отклоняемся от эталонной ситуации | R | = 1.

Тест сферичности Бартлетта

- **Статистический критерий** вычисляется по формуле:

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \times \ln|R|$$

- p – число переменных, n – число наблюдений (объектов).
- Статический критерий имеет распределение χ^2

  со степенью свободы ***df*** = p·(p-1)/2.
- Значимость критерия (p-value) можно найти с помощью функции (в R): **pchisq(chi2, df, lower.tail = F)**

chi2 – значение статистического критерия;
параметр **lower.tail**: если ИСТИНА (по умолчанию), вероятность P [X≤x], иначе P [X> x].
Функция рассчитывает p-значение. Например, если p-значение = 4,35 x 10^(-8) <0,05.
Нулевая гипотеза отклоняется на уровне 5%.
Можно эффективно применить PCA для исследуемого набора данных.

Тест сферичности Бартлетта

- Можно использовать встроенную фукцию:

```
> library(psych)
> print(cortest.bartlett(R,n = nrow(data)))
$chisq
[1] 54.25167

$p.value
[1] 4.355652e-08

$df
[1] 10
```

- R – корреляционная матрица
- data – исходные данные
- n – число наблюдений (объектов)

Тест сферичности Бартлетта

- *Примечание*: тест Бартлетта имеет недостаток. Когда количество объектов «n» увеличивается, он всегда оказывается статистически значимым.
- В некоторых источниках рекомендуется использовать этот тест, только если соотношение «n/p» (количество объектов, деленное на количество переменных) меньше 5.

Критерий КМО

Тест факторной адекватности КМО

(**Kaiser**-Meyer-Olkin).

(Кайзер, МВА – мера выборочной адекватности)

- **Индекс КМО, если индекс** ≈ 1 (от 0.6 до 1), метод PCA применим.
- **Индекс КМО, если индекс** ≈ 0, метод PCA не применим.
- **Основан на сравнении корреляций и *частных корреляций.***
- Прогнозирует, насколько хорошо факторизуются данные, на основе корреляции и частной корреляции.

Критерий КМО

- Отправной точкой является **корреляционная** матрица.
- Пусть переменные более или менее коррелированы, но на корреляцию между двумя переменными могут влиять другие переменные.
- При изучении многомерных связей парные корреляции могут давать совершенно неверные представления о характере связи между двумя переменными. Высокий коэффициент корреляции может быть обусловлен влиянием других переменных, как учтенных, так и неучтенных.
- Поэтому в многомерном случае предлагается использовать **матрицу частных корреляций**, чтобы измерить силу линейной связи между переменными, очищенную от влияния других факторов.

- Общий индекс КМО

$$KMO = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

- r_{ij} - наблюдаемые коэффициенты корреляции
- a_{ij} - частные коэффициенты корреляции

Матрица частных корреляций

- R - корреляционная матрица, $V=(v_{ij})$ - обратная к R.
- Матрица частных корреляций $A=(a_{ij})$. Может быть получена из матрицы корреляции:

$$a_{ij} = -\frac{v_{ij}}{\sqrt{v_{ii} \times v_{jj}}}$$

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2m} \\ r_{31} & r_{32} & 1 & \dots & r_{3m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & r_{m3} & \dots & 1 \end{bmatrix}, \quad V = R^{-1} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ v_{m1} & v_{m2} & \dots & v_{mm} \end{bmatrix}$$

- Частные коэффициенты корреляции характеризуют взаимосвязь между двумя выбранными переменными при исключении влияния остальных показателей

$$KMO = \frac{\sum_{i}\sum_{j\neq i} r_{ij}^2}{\sum_{i}\sum_{j\neq i} r_{ij}^2 + \sum_{i}\sum_{j\neq i} a_{ij}^2}$$

- При вычислении общего КМО в числитель вносится сумма квадратов корреляций всех переменных в данном анализе (за исключением корреляций переменных с самими собой 1.0, конечно).
- В знаменатель вносится та же самая сумма плюс сумма квадратов частных корреляций каждой переменной *i* с каждой переменной *j*, контролирующих другие связи в этом анализе.
- Идея состоит в том, что частная корреляция не должна быть слишком высокой, если ожидается, что в результате факторного анализа должны возникать отличающиеся факторы.
- **Если частная корреляция близка к нулю, PCA может эффективно выполнить факторизацию, потому что переменные сильно связаны: КМО≈1.**

Частный индекс КМО

- Общий индекс КМО

$$KMO = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

- Частный индекс КМО

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

- Мы можем вычислить индекс КМО для каждой переменной, чтобы обнаружить те, которые не связаны с другими переменными.

Population	School	Employment	Services	HouseValue
0.47207897	0.55158839	0.48851137	0.80664365	0.61281377

Частные индексы КМО:

- Рассмотрим Население (0,472) и Занятость (0,488). Они сильно коррелированы между собой (r = 0,9724), но не коррелированы с другими переменными.
- Частные корреляции

	Population	School	Employment	Services	HouseValue
Population	1.00000	-0.54465	0.97083	0.09612	0.15871
School	-0.54465	1.00000	0.54373	0.04996	0.64717
Employment	0.97083	0.54373	1.00000	0.06689	-0.25572
Services	0.09612	0.04996	0.06689	1.00000	0.59415
HouseValue	0.15871	0.64717	-0.25572	0.59415	1.00000

- Эти переменные не связаны с другими (которые определяют первый фактор), они определяют второй фактор PCA.

Рекомендация

- Значение КМО варьируется от 0 до 1.0 и для проведения факторного анализа общий КМО должен составлять 0.6 или выше.
- Если это не так, рекомендуется отбрасывать индикаторы-переменные с наименьшими значениями индивидуальной статистики КМО до тех пор, пока общий КМО не достигнет 0.6. (Некоторые исследователи используют более мягкую отсечку 0.5.)

- Тест сферичности Бартлетта и индекс КМО позволяют оценить возможность использования PCA.

Диаграмма biplot

- Отображение переменных и наблюдений (объектов) в пространстве d (= 2 или 3) измерений.
- Применяется в PCA
- Наблюдения – точки.
- Переменные в виде вектора.
- Углы между векторами

показывают корреляции.

№	Populatio	School	Employme	Services	HouseValue
1	5700	12,8	2500	270	25000
2	1000	10,9	600	10	10000
3	3400	8,8	1000	10	9000
4	3800	13,6	1700	140	25000
5	4000	12,8	1600	140	25000
6	8200	8,3	2600	60	12000
7	1200	11,4	400	10	16000
8	9100	11,5	3300	60	14000
9	9900	12,5	3400	180	18000
10	9600	13,7	3600	390	25000
11	9600	9,6	3300	80	12000
12	9400	11,4	4000	100	13000