Метод главных компонент (МГК)
Principal components analysis (PCA)

Метод главных компонент родственник регрессионного анализа
Метод главных компонент родственник факторного анализа

МГК и SVD, заполнение пропусков

Предложен Pearson (1901) и независимо Hotelling (1933)

Проклятие размерности.
Не работает интуиция, многое меняется.
Длина диагонали единичного куба стремится к бесконечности.
Шар вписан в единичный куб. Объем стремится к
Объем шара радиуса r в R^k .

Базовая идея (для математиков)
имеем набор переменных
желаем получить новый набор переменных, чтобы они были

- некоррелированными
- линейными комбинациями исходных
- их общая дисперсия такая же

Имеем оценки студентов. Все студенты сдавали экзамены по одним и тем же дисциплинам.
Оценка каждого экзамена — отдельная переменная.
Надо построить индекс (одну переменную) того, как хорошо студент сдавал экзамены.

Идея 1. Сосчитать средний балл.
Идея 2. Сосчитать взвешенное среднее. Как выбрать веса?
Идея 3. ??? Применить метод главных компонент

Некоторые оценки надо предварительно стандартизировать. (Если преподаватель ставит только высокие или только низкие отметки)

Вопрос на будущее: у каких оценок будет максимальный вес, если применить метод главных компонент

Аналогичная задача: создание индексов в экономике
- индекс зарплат
- индекс цен на цветные металлы
- индекс цен на снимаемое жилье
- индекс стоимости жизни в разных городах

- и т.д.

Каждый раз индекс используется для сопоставления показателей в пространстве (например сравнение стоимости жизни в разных городах мира) или во времени (например изменения цен на цветные металлы)

Морфология (биология) — наука о форме и строении организмов.
Рыбы
1-я компонента — размер
2-я компонента — отношение высоты к длине

В некоторых приложениях главные компоненты являются целью анализа сами по себе, их интерпретируют, они измеряют что-то, не поддающееся измерению.
Чаще всего они получены для использования в качестве входных данных для другого анализа.

Результаты МГК могут применяться в регрессионном анализе в следующих ситуациях.
Если слишком много объясняющих переменных сравнительно с количеством наблюдений, то вместо части исходных переменных используем ГК (Вопрос: вместо какой части переменных?);
Если объясняющие переменные сильно коррелированы (проблема мультиколлинеарности), вместо них используем ГК, которые не коррелируют по построению. (Rencher 1995).

Анализ главных компонент. Математическое описание

Рассмотрим случайный вектор $X_1, X_2, \ldots, X_k$

Задача 1. Найти линейную комбинацию $Y_1 = a_{11} X_1 + a_{12} X_2 + \ldots + a_{1k} X_k$ такую, что $D(Y_1)$ максимальна.

Дополнительное условие 1: $\vec{a}_1 \cdot \vec{a}_1^{\,T} = 1$, где $\vec{a}_1 = (a_{11,} a_{12,} \ldots, a_{1k})$

Задача 2. Найти линейную комбинацию $Y_2 = a_{21} X_1 + a_{22} X_2 + \ldots + a_{2k} X_k$ такую, что $D(Y_2)$ максимальна.

Дополнительное условие 2.1: $\vec{a}_2 \cdot \vec{a}_2^{\,T} = 1$, где $\vec{a}_2 = (a_{21,} a_{22,} \ldots, a_{2k})$

Дополнительное условие 2.2: $corr(Y_{2,} Y_1) = 0$

Задача 3. Найти линейную комбинацию $Y_3 = a_{31} X_1 + a_{32} X_2 + \ldots + a_{3k} X_k$ такую, что $D(Y_3)$ максимальна.

Дополнительное условие 3.1: $\vec{a}_3 \cdot \vec{a}_3^{\,T} = 1$, где $\vec{a}_3 = (a_{31,} a_{32,} \ldots, a_{3k})$

Дополнительное условие 3.2: $corr(Y_{3,} Y_1)=0$
Дополнительное условие 3.3: $corr(Y_{3,} Y_2)=0$

Задача k. Найти линейную комбинацию $Y_k = a_{k1} X_1 + a_{k2} X_2 + ... + a_{kk} X_k$ такую, что $D(Y_k)$ максимальна.
Дополнительное условие k.1: $\vec{a}_k \cdot \vec{a}_k{}^T = 1$, где $\vec{a}_k = (a_{k1,} a_{k2,} ..., a_{kk})$
Дополнительное условие k.2: $corr(Y_k, Y_1)=0$
Дополнительное условие k.3: $corr(Y_k, Y_2)=0$
…
Дополнительное условие k.k: $corr(Y_k, Y_{k-1})=0$

Обозначение. ***R*** **–** матрица ковариаций (корреляций) случайного вектора ***X***.

Решение задач 1-k

Каждая из ***k*** задач нахождения вектора $\vec{a}_i$ $i=1,2,...,k$ сводится к решению одного и того же уравнения. Доказательство см Кендалл, Стьюарт "Многомерный статистический анализ и временные ряды" стр 400 .

$$R \cdot a^T = \lambda \cdot a^T$$

Это задача нахождения собственных чисел и собственных векторов матрицы ковариаций (корреляций) ***R.*** Вспомним линейную алгебру. Надо решить систему

$$R \cdot a_i = \lambda_i \cdot a_i$$

Часто для этого сначала ищут корни многочлена

$$|R - \lambda \cdot E| = 0$$

Кроме того, верно, что $D(Y_i) = \lambda$
Решить недостаточно, разберемся с неочевидными (для нематематика) деталями.

Замечание 1. Параметр λ - собственное число матрицы R. Таких собственных чисел k, столько же, сколько случайных величин.

Все собственное числа вещественные и не отрицательные. Опираемся на следующие три утверждения.

Утверждение 1. Если вещественная квадратная матрица неотрицательно определенная, то ее собственные числа вещественные и не отрицательные.

Утверждение 2. Ковариационная матрица неотрицательно определенная.

Утверждение 3. Корреляционная матрица является ковариационной матрицей

Определение стандартизации

Будем говорить, что провели стандартизацию случайного вектора $X=(X_1, X_2, \ldots, X_k)$, если создали новый случайный вектор ***Y***, координаты которого получены по формуле $Y_i=\dfrac{X_i-EX_i}{DX_i}$

Утверждение 4. Корреляционная матрица вектора X равна ковариационной матрице вектора Y

Замечание 2. Дисперсия равна собственному числу. Поэтому в литературе дисперсия часто называется собственным числом, а не дисперсией, что путает начинающих.

Замечание 3. В задаче 1 решением будет собственный вектор, соответствующий максимальному собственному числу. В задаче i - собственный вектор, соответствующий i-ому по величине собственному числу.

Замечание 4. Известно, что если из собственных векторов составить матрицу, она будет ортогональной. Обозначим эту матрицу ***A***. Справедливо равенство

$$A^{-1}\cdot R\cdot A=\Lambda$$

где

$$\Lambda=\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}$$

Замечание 5.

$$\sum_{i=1}^{k}\lambda_i=trace(R)$$

Для доказательства используем равенство $A^{-1}\cdot R\cdot A=\Lambda$ и свойство $trace(A\cdot B)=trace(B\cdot A)$

Методы определения числа факторов (Для случая корреляционной матрицы).

1. Оставить те главные компоненты, для которых собственное число больше 1 (Kaiser, Kaiser – Guttman)
2. Оставить те главные компоненты, для которых собственное число больше 0.8 (Jolliffe)
3. Подобрать ***p*** - число главных компонент так, чтобы отношение

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

превысило 0.8 (или 0.7 ...)
4. Найти точку излома графика каменистая осыпь (scree plot, elbow plot, Cattell)

5. Minka's MLE — но готовы ли Вы согласиться с предположением, что вектор переменных имеет нормальное распределение?
(Minka, Automatic choice of dimensionality for PCA, 2000)
6. Karlis – Saporta - Spinaki rule

$$\lambda > 1 + 2 \cdot \sqrt{\frac{k-1}{n-1}}$$

7. …

Стоит ли проводить факторный анализ?

Теста Bartlett'a

H_0 : Корреляционная матрица совпадает с единичной.

More useful for pedagogical purposes than actual applications.
(Из мануала к пакету psych в R
https://personality-project.org/r/html/cortest.bartlett.html)

Статистика критерия: -ln(det(R) * (N-1 — (2p+5)/6)

Статистика имеет распределение хи-квадрат, если корреляционная матрица R единичная.

Если p-значение мало (например меньше, чем 0.05), то факторный анализ данных может быть полезным.

Используется также для проверки того, что все корреляции остатков равны нулю.

В пакете R реализован в библиотеке psych
Процедура `cortest.bartlett()`

В Python популярна обертка для библиотеки psych

Используют несколько других тестов Бартлета.
Например   Bartlett's Test for Equality of Variances

Правильная ссылка
Bartlett, M. S.,
The Effect of Standardization on a chi square Approximation in Factor Analysis,
(1951), Biometrika, 38, 337-344.

Статистики Kaiser-Meyer-Olkin

Традиционное название **KMO** (Kaiser-Meyer-Olkin)

Формула для статистики критерия приведена в
https://www.rdocumentation.org/packages/REdaS/versions/0.9.3/topics/Kaiser-Meyer-Olkin-Statistics

Процедура выдает
А). общую оценку для всего набора переменных,
Б). индивидуальные оценки для каждой переменной.

Часто случается, что общая оценка высокая, но при этом некоторые индивидуальные оценки низкие. Любая переменная с оценкой ниже 0,5 плохо описывается моделью, ее можно отнести к индивидуальным факторам. Из факторного анализа ее стоит исключить, а тест нужно провести заново.

Интерпретация значений по Кайзеру

Величина оценки	Интерпретация значения
≥.9	marvelous
[.8, .9)	meritorious
[.7, .8)	middling
[.6, .7)	mediocre
[.5, .6)	miserable
<.5	unacceptable

Ссылки

Kaiser, H. F. (1970). A Second Generation Little Jiffy. Psychometrika, 35(4), 401--415.

Kaiser, H. F. (1974). An Index of Factorial Simplicity. Psychometrika, 39(1), 31--36.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. Educational and Psychological Measurement, 34, 111--117.