

# Робастные методы сжатия данных

Парамонов Всеволод Антонович

2024

# Цели

- Изучение метода главных компонент и робастных оценок корреляционных матриц
- Проведение компьютерного моделирования, имитирующего коррелированные данные с различными типами засорения
- Подбор больших публичных датасетов для анализа сжатия
- Сжатие данных с помощью метода главных компонент
- Построение метрик, оценивающих качество сжатия данных

# Метод главных компонент

## Понижение размерности

- процесс сокращения количества измерений (признаков) с сохранением наиболее значимой информации о структуре данных

$$f : \mathbb{R}^d \rightarrow \mathbb{R}^k, \quad k < d$$

# Мотивы понижения размерности

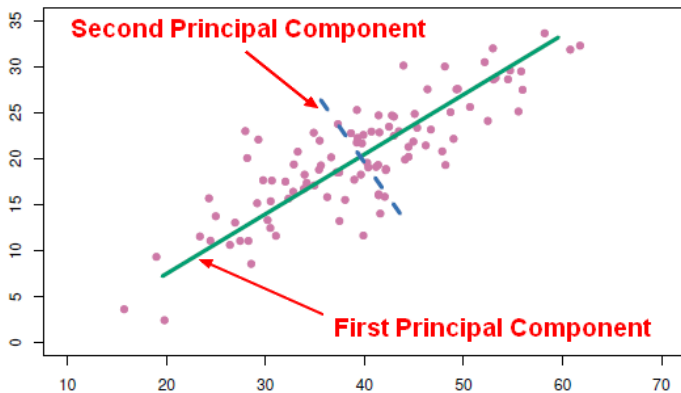
- Сжатие данных для более эффективного хранения информации
- Визуализация и интерпретация данных
- Уменьшение вычислительных затрат (например, в задачах регрессии)
- Борьба с мультиколлинеарностью

# Метод главных компонент

## МГК (РСА)

- линейное преобразование исходных признаков, направленная на поиск новых ортогональных осей таким образом, чтобы проекция исходных данных на эти оси имела максимальную дисперсию

# Метод главных компонент



# Метод главных компонент

Пусть  $X = [X_1, X_2, \dots, X_k] \in \mathbb{R}^{n \times k}$  - матрица объекты-признаки, где  $X_i \in \mathbb{R}^n$ . Главной задачей является переход от матрицы  $X$  к матрице  $Z \in \mathbb{R}^{n \times m}$ , причем  $m < k$ . Далее будем считать, что данные в матрице  $X$  центрированы и стандартизированы

Также необходима восстанавливаемость исходной матрицы признаков с приемлемой точностью. Таким образом  $\exists U \in \mathbb{R}^{k \times m}$  ( $U$  - ортонормированная матрица,  $U^T U = U U^T = I$ ), что матрица  $Z U^T$  является неплохим приближением матрицы  $X$



# МГК. Постановка задачи

## Максимизация дисперсии

Предположим, что признак  $Z_1$  линейно выражается через все признаки матрицы  $X$ . Тогда:

$$Z_1 = w_{11}X_1 + \dots w_{1k}X_k = Xw_1, \quad \sum_{i=1}^k w_{1i}^2 = 1$$

Тогда задача выглядит следующим образом:

$$\begin{cases} \|Z_1\|^2 = w_1^T X^T X w_1 \rightarrow \max_{w_1} \\ w_1^T w_1 = 1 \end{cases}$$

# МГК. Постановка задачи

Максимизация дисперсии

Выпишем Лагранжиан:

$$\mathbb{L} = w^T X^T X w - \lambda(w^T w - 1) \rightarrow \max$$

$$\frac{\partial \mathbb{L}}{\partial w} = 2X^T X w - 2\lambda w = 0$$

$$X^T X w = \lambda w$$

Подставляя решение в функцию Лагранжа получается:

$$\mathbb{L} = \lambda w^T w - \lambda(w^T w - 1) = \lambda$$

**Итого:** максимальное значение дисперсии компоненты достигается при использовании собственного вектора, которому соответствует максимальное собственное значение, в качестве компоненты

Допустим, что пока что не происходит сокращения размерности, тогда:

$$Z = XU \Rightarrow X = ZU^T$$

Разложим матрицу  $X$  с помощью SVD:

$$X = \underbrace{VD}_Z U^T = ZU^T$$

Тогда матрица ковариаций для главных компонент будет выглядеть следующим образом:

$$Z^T Z = U^T X^T X U = U^T U D^T V^T V D U^T U = D^T D$$

Получилось так, что ковариационная матрица  $Z^T Z$  - диагональная матрица, у которой компоненты не коррелированы, а на диагонали находятся дисперсии данных компонент

Теперь представим, что мы берем не все компоненты, а лишь  $m$  штук. Тогда ошибка будет выглядеть следующим образом:

$$\epsilon = \|XU - X\hat{U}\|_F^2,$$

где  $\hat{U}$  - матрица, у которой занулены столбцы, начиная с  $m + 1$ . Домножим выражение на  $U^T$ , поскольку норма Фробениуса не зависит от домножения на ортогональную матрицу

$$\epsilon = \|XUU^T - X\hat{U}U^T\|_F^2$$

Разложим матрицу  $X$  через SVD и подставим в выражение:

$$\epsilon = \|VDU^T - V\hat{D}U^T\|_F^2 = \|V(D - \hat{D})U^T\|_F^2$$

$D$  и  $\hat{D}$  - диагональные матрицы  $\Rightarrow$  на диагонали матрицы  $(D - \hat{D}) = D^*$  находятся элементы  $(0, \dots, \sqrt{\lambda_{m+1}}, \dots)$ . Тогда:

$$\epsilon = \|UD^*V^T\|_F^2 = \text{tr}(UD^*V^TVD^{*T}U) = \text{tr}(D^2) = \sum_{i=m+1}^k \lambda_i$$

**Итого:** Ошибка будет соответствовать сумме квадратов собственных значений, которым соответствуют не использовавшиеся компоненты

# Недостатки МГК

- Сложность вычисления главных компонент при большом количестве признаков. Данная проблема решается **степенным методом** нахождения собственных значений и собственных векторов
- МГК способен находить только линейные подпространства исходного пространства, которые сохраняют дисперсию исходных данных с высокой точностью. Таким образом от обычного МГК можно перейти к **ядровому МГК**
- МГК чувствителен к выбросам данных и зависит от стандартизации

# Ядровый МГК

Идея ядрового МГК заключается в применении нелинейных преобразований к векторам признаков, которое сделало бы линейные методы более мощными

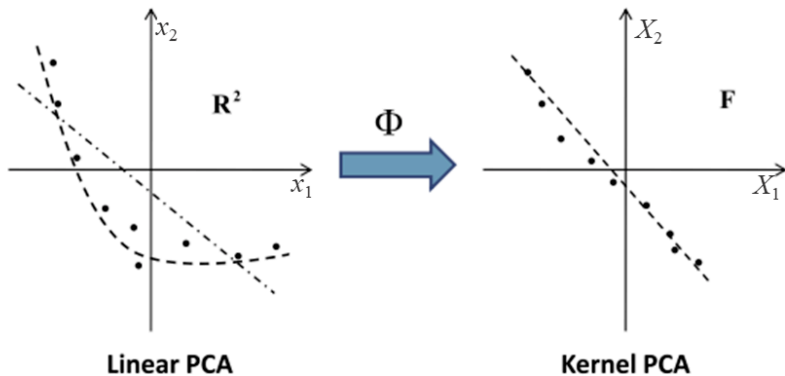
Вместо обычного скалярного произведения  $\langle x, y \rangle$  в пространстве  $\mathbb{R}^n$  в обычном МГК теперь используется следующее скалярное произведение:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle,$$

где  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^N$ ,  $N \gg n$ . Также для применения функции  $\phi$  нет необходимости знать явный вид функции



# Ядровый МГК



# Ядровый МГК

В качестве функции  $\phi$  можно использовать следующие функции / ядра:

- Линейное ядро:  $K(x, y) = xy$
- Гауссово ядро:  $K(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$
- Полиномиальное ядро:  $K(x, y) = (\gamma xy + c)^d$
- ...

# Выбор количества компонент

## Установление порога

Дисперсиями главных компонент являются соответствующие собственные значения ковариационной матрицы стандартизированных величин, упорядоченные по убыванию. Тогда общая доля дисперсии, которую покрывают  $l$  компонент:

$$\frac{\sum_{i=1}^l \lambda_{(i)}}{\sum_{j=1}^k \lambda_j}$$

После того как был установлен порог  $t$  (удовлетворительная доля дисперсии, покрываемая главными компонентами), подбирается такое  $l$ , чтобы значение общей доли было больше порога

# Выбор количества компонент

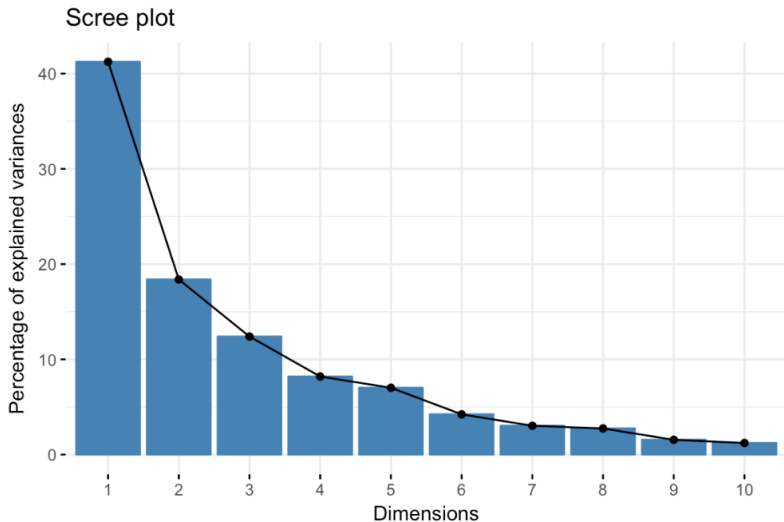
## Критерий Кайзера

В соответствии с правилом Кайзера, следует выбирать те компоненты, у которых собственное значение больше 1

$$\Lambda = [c_i, c_{i+1}, \dots] \forall \lambda_i \geq 1$$

Где  $\Lambda$  - матрица с выбранными главными компонентами

# Выбор количества компонент



# Оценка качества сжатия данных

Для сравнения качества между моделями сжатия многомерных данных используется показатель:

$$e = \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^m \hat{\lambda}_i} - 1,$$

где  $\lambda$  - собственное значение истинной корреляционной матрицы,  
 $\hat{\lambda}$  - собственное значение оцененной корреляционной матрицы

Идеальным значением показателя будет 0, поэтому в дальнейшем можно рассматривать модуль данного показателя

# Робастные оценки корреляционных матриц

# Корреляция Пирсона

$$\rho(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}}$$



# Корреляция Спирмена

- Для каждого значения признака присваивается ранг от наименьшего к наибольшему. Если же встречаются повторяющиеся признаки, то им присваивается среднее значение ранга
- Для каждой пары значений признаков вычисляется квадрат разности рангов

$$d_i = (\text{rang}_i^{x_j} - \text{rang}_i^{x_k})^2$$

Где  $j, k$  - номера признаков,  $i$  - номер значений признаков

- Расчет значения корреляции

$$\rho(d_i) = 1 - \frac{6 \sum_{i=1}^n d_i}{n(n^2 - 1)}$$

# Медианная корреляция

$Med_X, Med_Y$  – медианы  $X$  и  $Y$  соответственно

$$MAD(X) = Med(|X_i - Med(X_i)|)$$

$$MAD(Y) = Med(|Y_i - Med(Y_i)|)$$

$$Covmed(X, Y) = [(X - Med_X)(Y - Med_Y)]$$

$$\rho(X, Y) = \frac{Covmed(X, Y)}{MAD(X) \times MAD(Y)}$$

$$\rho(u, v) = \frac{(MAD^2(u) - MAD^2(v))}{(MAD^2(u) + MAD^2(v))}$$

$$u = \frac{X - Med(X)}{\sqrt{2}MAD(X)} + \frac{Y - Med(Y)}{\sqrt{2}MAD(Y)}$$

$$v = \frac{X - Med(X)}{\sqrt{2}MAD(X)} - \frac{Y - Med(Y)}{\sqrt{2}MAD(Y)}$$

# Работа с синтетическими данными

# Генерация данных

Сгенерируем данные из следующих распределений для проведения экспериментов с целью оценки качества робастности оценок корреляционных матриц и сжатия многомерных данных:

- Выборка 1: многомерное нормальное распределение

$$(X)_{i=1}^n \sim \mathbb{N}(0, K)$$

- Выборка 2: многомерное распределение Тьюки

$$(X)_{i=1}^n \sim (1 - \delta)\mathbb{N}(0, K) + \delta\mathbb{N}(0, c^2 K)$$

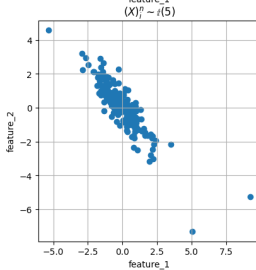
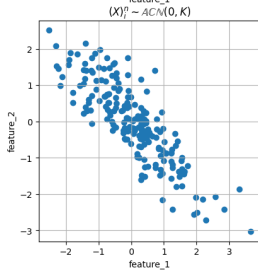
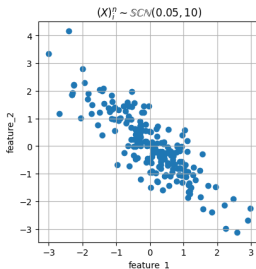
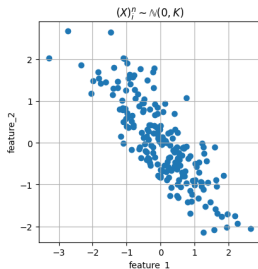
- Выборка 3: многомерное распределение ACN

$$(X)_{i=1}^n \sim (1 - \delta)\mathbb{N}(0, K) + \delta\mathbb{N}(\mu, c^2 K)$$

- Выборка 4: многомерное распределение Стьюдента

$$(X)_{i=1}^n \sim t(df)$$

# Визуализация данных



# Оценка робастности корреляционных функций

Для оценки робастности  $k$  раз сгенерируем выборку из некоторого распределения и оценим корреляционную матрицу. В качестве метрики схожести корреляционных матриц (истинной и сгенерированной) будем использовать евклидово расстояние между векторами, представляющими верхний треугольник (без диагональных элементов) матриц корреляции

Итоговой оценкой робастности будет усредненное значение евклидова расстояния по  $k$  генерациям

# Сравнение результатов оценивания для корреляционных функций

	Pearson	Spearman	MAD	MED
MultiNormal	0.208439	0.212653	5.219468	2.047972
ACN	0.206961	0.204515	9.596546	2.046548
SCN	0.191703	0.202676	9.682890	2.045978
t	0.283220	0.236937	5.505540	2.045801

Таблица: Средние "ошибки" оценок корреляционных матриц



# Оценка разброса значений корреляционных функций

Сгенерируем для каждого распределения  $k$  выборок и для каждой из них посчитаем корреляционную матрицу. Соберем в матрицу все получившиеся оценки, где в каждой строке будут находиться значения верхнего треугольника корреляционной матрицы.

Далее посчитаем стандартное отклонение каждого столбца матрицы и посчитаем квадрат евклидовой нормы получившегося вектора, что и будет являться мерой разброса корреляционной функции

# Проверка робастности корреляционных функций

	Pearson	Spearman	MAD	MED
MultiNormal	0.043312	0.05319	0.037553	0.052439
ACN	0.046013	0.043703	0.043022	0.041592
SCN	0.047638	0.045243	0.049953	0.040631
t	0.142757	0.101361	0.116142	0.105153

Таблица: Дисперсии оценок корреляционных матриц

# Итоги. Выбор корреляционной функции

Показатели отражают, что следует выбрать корреляцию Пирсона в качестве функции для оценки корреляционных матриц, поскольку данная функция имеет маленький разброс и ошибку относительно других функций оценки корреляции

Также стоит отметить, что корреляция Спирмена составила неплохую конкуренцию, тк ее показатели не сильно отличаются от показателей корреляции Пирсона

# Сравнение МГК с разными корреляционными функциями

По аналогии проведем несколько экспериментов, где для каждого набора данных и для каждой корреляционной функции посчитаем главные компоненты и возьмем первые 3 компоненты и посчитаем разброс проекций исходных данных на эти компоненты и показатель для сравнения моделей МГК (слайд 22):

$$e = \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^m \hat{\lambda}_i} - 1$$

# Сравнение дисперсий компонент

	Pearson	Spearman	MAD	MED
MultiNormal	1.860765	1.860087	0.248955	1.037364
ACN	1.876917	1.876182	0.243319	0.229604
SCN	1.872717	1.871476	0.241794	0.886149
t	1.89755	1.8936	0.184575	1.270027

Таблица: Дисперсии главных компонент

# Сравнение моделей по показателю близости истинных и оцененных собственных значений корреляционной матрицы

	Pearson	Spearman	MAD	MED
MultiNormal	0.004223	0.005806	0.321503	145.589495
ACN	0.004419	0.005942	0.363547	145.134802
SCN	0.002186	0.004058	0.379665	101.694736
t	0.015245	0.005533	0.779886	119.448396

Таблица: Дисперсии главных компонент

# Выводы. МГК

Было показано, что действительно качество сжатия многомерных данных зависит от степени робастности выбранной корреляционной функции. Также были подтверждены выводы из предыдущего пункта о робастности корреляций Пирсона и Спирмена: модели МГК с использованием данных корреляционных функций показали наилучшие результаты

# Работа с реальными данными