Доказательство: Главные компоненты совпадают с собственными векторами матрицы ковариаций

1. Формулировка задачи PCA

Задача PCA заключается в нахождении направлений, которые максимально сохраняют дисперсию данных. Эти направления называются **главными компонентами** и представляют собой линейные комбинации признаков, вдоль которых дисперсия проекций данных максимальна.

Пусть $X \in \mathbb{R}^{n \times m}$ — матрица данных, где n — количество объектов, а m — количество признаков. Сначала данные центрируются, после чего вычисляется ковариационная матрица:

$$\Sigma = \frac{1}{n-1} X^\top X$$

Матрица Σ содержит информацию о дисперсиях и ковариациях между признаками.

2. Задача максимизации дисперсии

Требуется найти такое направление $v \in \mathbb{R}^m$, вдоль которого дисперсия проекций максимальна. Проекция данных на v:

$$Y = Xv$$

Дисперсия этой проекции:

$$\mathrm{Var}(Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i^\top v)^2$$

Это выражение можно записать в матричной форме:

$$\mathrm{Var}(Y) = v^\top \Sigma v$$

Задача PCA: найти вектор v, максимизирующий $v^\top \Sigma v$ при условии $\|v\| = 1$.

3. Свойства матрицы ковариаций

Матрица Σ симметрична и положительно полуопределённая. Это означает, что все её собственные значения вещественные, а собственные векторы — ортогональны.

Согласно теореме о максимуме квадратичной формы, максимальное значение $v^\top \Sigma v$ при $\|v\| = 1$ достигается тогда и только тогда, когда v является собственным вектором Σ, соответствующим наибольшему собственному значению.

4. Собственные значения и собственные векторы

Задача:

$$\max_v v^\top \Sigma v \quad \text{при условии } \|v\| = 1$$

Решается с помощью метода множителей Лагранжа. Вводим лагранжиан:

$$\mathcal{L}(v, \lambda) = v^\top \Sigma v - \lambda(v^\top v - 1)$$

Найдем производную по v и приравняем к нулю:

$$\frac{\partial \mathcal{L}}{\partial v} = 2\Sigma v - 2\lambda v = 0 \quad \Rightarrow \quad \Sigma v = \lambda v$$

Это уравнение показывает, что v должен быть собственным вектором Σ, а λ — соответствующим собственным значением.

5. Вывод

таким образом, оптимальные направления PCA — это собственные векторы матрицы ковариаций Σ, а дисперсия вдоль этих направлений равна соответствующим собственным значениям. Именно поэтому главные компоненты совпадают с собственными векторами Σ.