Метод главных компонент(principal component analysis) - Метод понижения размерности выборки, с наименьшей потерей информации. Идея метода коррелирует с LS решением (linear regression), поскольку главная компонента(ось) совпадает с линейной аппроксимацией выборки.

Figure 1: Двумерная выборка и линейная регрессионная модель

Цель этого метода заключается в поиске таких осей, на которые можно спроектировать многомерную выборку и сохранить при этом наибольшее количество информации. Сравним два варинта. Проекция выборки на оптимальную линейную поверхность аппроксимирующую выборку и на произвольную прямую:

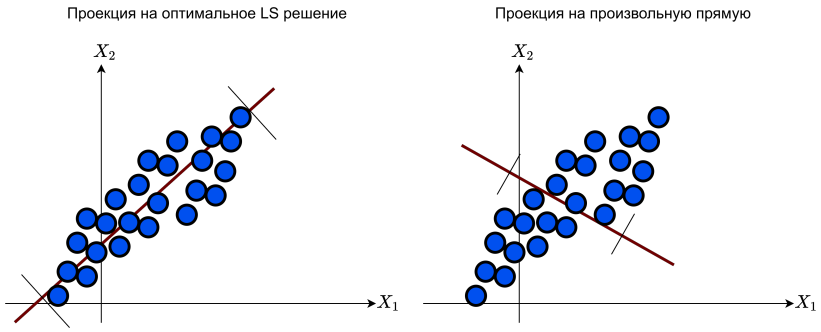

Figure 2: Сравнительное представление двух вариантов проектирования выборки на разные прямые

С точки зрения теории информации - степень неопределенности характеризует информативность данных. Известно, что дисперсия данных определяет их квадрат отклонения от среднего или же с точки зрения теории информации это понятие коррелирует с неопределенностью. Таким образом, чтобы уменьшить размерность выборки, сохраняя ее информативность, необходимо найти такое афинное множество, проектируя на которое данные, их дисперсия будет максимальна. Сформулируем критерий.

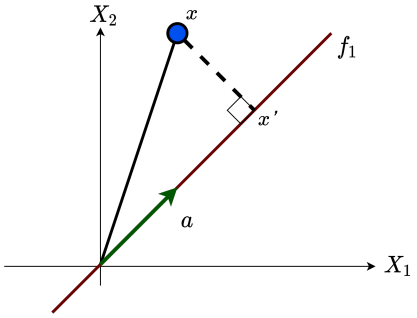

Figure 3: Геометрческая предпосылка к формулированию критерия

Положим, что $\boldsymbol{a} = (a_1, a_2, ..., a_n)^T$- конечномерный базис линейного множества, на которое планируется спроектировать выборку, где $\|\boldsymbol{a}\| = 1$. Естественным образом будет думать , что длина спроектированного вектора $\|x_i'\|$ это взвешенный базис $\boldsymbol{a} = (a_1, a_2, ..., a_n)^T$.

$$\|x_i'\| = \frac{\overline{\boldsymbol{a}}}{\|\boldsymbol{a}\|}\overline{x_i} = \|x_i\| cos\theta = proj_{\boldsymbol{a}} x_i = \boldsymbol{a^T} x_i$$

Поскольку цель - найти прямую с максимальной дисперсией проекций - это задача условной оптимизации:

$$\begin{cases} var\left(\boldsymbol{a^T} x_i - \overline{x}\right) \to \max\limits_{a} \\ \|\boldsymbol{a}\| - 1 = 0 \end{cases}$$

Положим, что данные смещены к центру на координату МО, поэтому задача упрощается до:

$$\begin{cases} \frac{1}{n}\sum_i^n \left(\boldsymbol{a^T} x_i\right)^2 \to \max\limits_{a} \\ \|\boldsymbol{a}\| - 1 = 0 \end{cases}$$

Раскладывая $\frac{1}{n}\left[(a_1x_1)^2 + (a_2x_2)^2 + ... + (a_nx_n)^2\right] = \frac{1}{n}\left\langle \begin{array}{cc} \boldsymbol{a^T}X & \left(\boldsymbol{a^T}x\right)^T \end{array} \right\rangle = \frac{1}{n}\boldsymbol{a^T}X\left(\boldsymbol{a^T}X\right)^T = \boldsymbol{a^T}\frac{XX^T}{n}\boldsymbol{a} = \boldsymbol{a^T}S\boldsymbol{a}$ и по свойству: $\|\boldsymbol{a}\| = \boldsymbol{a^T}\boldsymbol{a}$ задача трансформируется к:

$$\begin{cases} \boldsymbol{a^T}S\boldsymbol{a} \to \max\limits_{a} \\ \boldsymbol{a^T}\boldsymbol{a} - 1 = 0 \end{cases} \tag{1}$$

Задачи условной оптимизации решаются методом множителей Лагранжа:(ищем $\boldsymbol{a}$=> дифференцируем по $\boldsymbol{a}$)

$$L(\boldsymbol{a}, S, \lambda) = \boldsymbol{a^T}S\boldsymbol{a} - \lambda\left(\boldsymbol{a^T}\boldsymbol{a} - 1\right)$$

$$\frac{dL(\boldsymbol{a}, S, \lambda)}{\partial \boldsymbol{a}} = \boldsymbol{a^T}S\boldsymbol{a} - \lambda\left(\boldsymbol{a^T}\boldsymbol{a} - 1\right)$$

$$\frac{dL(\boldsymbol{a}, S)}{d\boldsymbol{a}} = \langle \boldsymbol{a}, S\boldsymbol{a}\rangle - \lambda\langle \boldsymbol{a}, \boldsymbol{a}\rangle + 1$$

$$\frac{dL(\boldsymbol{a}, \boldsymbol{S}, \boldsymbol{\lambda})}{d\boldsymbol{a}} = \langle \boldsymbol{a}, S\boldsymbol{a}\rangle - \lambda\langle \boldsymbol{a}, \partial\boldsymbol{a}\rangle + 1$$

$$dL(\boldsymbol{a}, S, \lambda) = d\langle \boldsymbol{a}, S\boldsymbol{a}\rangle - \lambda d\langle \boldsymbol{a}, \boldsymbol{a}\rangle$$

$$dL(\boldsymbol{a}, S, \lambda) = \langle d\boldsymbol{a}, S\boldsymbol{a}\rangle + \langle \boldsymbol{a}, dS\boldsymbol{a}\rangle - \lambda\langle d\boldsymbol{a}, \boldsymbol{a}\rangle - \lambda\langle \boldsymbol{a}, d\boldsymbol{a}\rangle$$

$$dL(\boldsymbol{a}, \boldsymbol{S}, \boldsymbol{\lambda}) = \langle S\boldsymbol{a}, d\boldsymbol{a}\rangle + \left\langle S^T\boldsymbol{a}, d\boldsymbol{a}\right\rangle - \lambda\langle \boldsymbol{a}, d\boldsymbol{a}\rangle - \lambda\langle \boldsymbol{a}, d\boldsymbol{a}\rangle$$

$$dL(\boldsymbol{a}, S, \lambda) = S\boldsymbol{a} + S^T\boldsymbol{a} - 2\lambda\boldsymbol{a}$$

Поскольку матрица квадратичной формы симметрична: $S = S^T$, то приравнивая дифференциал к 0, получим, что:

$$2S\boldsymbol{a^T} - 2\lambda\boldsymbol{a} = 0$$

$$(S - \lambda I)\,\boldsymbol{a} = 0 \tag{2}$$

Отсюда видно, что $\boldsymbol{a}$ - собственный вектор и λ - собственные числа. Решая систему уравнений, и извлекая максимальное собственное значение (определяет длину вектора) мы можем найти собственный вектор согласно критерию. Таким образом, найден вектор, на который можно спроектировать данные и сформировать сжатие с максимальным сохранением информации. Другим вопросом является то, что получен единственный вектор. Пусть этот вектор будет обозначаться как $\boldsymbol{a} = \boldsymbol{a_1}$**.** Если мы планируем спроектировать данные на плоскость, нам необходимо решить другую задачу оптимизации:

$$\begin{cases} \boldsymbol{a_2^T} S \boldsymbol{a_2} \to \max\limits_{\boldsymbol{a_2}} \\ \boldsymbol{a_1^T a_2 = 0} \\ \|\boldsymbol{a_2}\| - \boldsymbol{1 = 0} \end{cases}$$

Известно, что результат этого решения: $\boldsymbol{a_2}$, (который, кстати соответствует второму по величине собственному вектору из задачи (1)) ставля новую задачу оптимизации с новыми условями, включающими все предыдущие результаты, мы можем произвольно уменьшать/увеличивать размерность выборки. Более простым пояснением является то, что собственные векторы симметической матрицы являются ортогональными. Поэтому решить задачу можно и в одну итерацию.

Поиск решения разложением SVD

Для симметрических матриц сещствует взаимосвязь между собственными значениями и сингулярными значениями SVD разложения. Известно, что SVD раскладывает матрицу по закону:

$$X = U\Sigma V^H \tag{3}$$

Где

U - самосопряженный оператор вращения ($UU^T = I$)

Σ - оператор растяжения (Диагональная матрица)

$V^H = V^T$ - самосопряженный оператор вращения для $X \in R^{nxp}$ (Эрмитово сопряжение к V) и ($VV^T = I$)

Согласно (3), имеем $X^T = \left(U\Sigma V^T\right)^T = V\left(U\Sigma\right)^T = V\Sigma^T U^T = V\Sigma U^T$

$S = X^T X = V\Sigma U^T U\Sigma V^T = V\Sigma I\Sigma V^T = V\Sigma^2 V^T$

$X^T XV = SV = V\Sigma^2$ (по свойству самосопряженности: $VV^T = I$)

Сравнивая (2) $S\boldsymbol{a} = \lambda\boldsymbol{a}$ и $SV = V\Sigma^2$, видим, что $\boldsymbol{a} = \Sigma^2$, то есть:

$$\begin{pmatrix} \lambda_1 & \cdots & \cdots & 0 \\ \vdots & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \cdots & \cdots & 0 \\ \vdots & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix} \tag{4}$$

Таким образом, формулировка алгоритма метода главных компонент следующая:

Algorithm 1 Метод главных компонент

- Нормировать данные по среднему
- Рассчитать корреляционную матрицу для данных: $S = \frac{1}{N} X^T X$
- Определить собственные значения и собственные вектора корреляционной матрицы (можно разложить по собственным векторам или использовать SVD)
- Отсортируйте собственные вектора в порядке убывания, выберите требуемое количество и спроектируйте данные на новые вектора $\boldsymbol{a^T} x_i \forall i = \overline{1, n}$

Вывод:

Был рассмотрен метод главных компонент. Его целью является уменьшить размерность выборки, с максимальным сохранением информации. Это позволяет визуализировать многомерные данные, энтропийно кодировать их и так далее. Было показано то, что наилучшими плоскостями, на которые проектируется выборка - те плоскости, на которых данные будут иметь максимальную дисперсию. Поставлена и решена задача оптимизации, а также показана связь между собственными числами симметрической матрицы и оператором растяжения SVD разложения.