

# Методы PCA и SVD

Р. М. Селевенко

Московский Государственный Университет, Факультет ВМК

При обработке больших массивов данных представленных в виде таблицы объектов и признаков возникает потребность в обработке этой таблицы с целью упрощения структуры данных. А именно, выделения объектов представляющих собой частные случаи вместо части общей выборки. А также выделения признаков представляющих собой линейную комбинацию других признаков. Например, в случае предсказания цены на комнаты в гостинице, нам могут быть даны такие фичи как площадь комнаты, длина и ширина периметра комнаты. В этом случае фичи длины и ширины нам не нужны т.к. они являются простой линейной комбинацией площади. Также существует так называемое "проклятие размерности при увеличении размерности пространства количество данных возрастает экспоненциально.

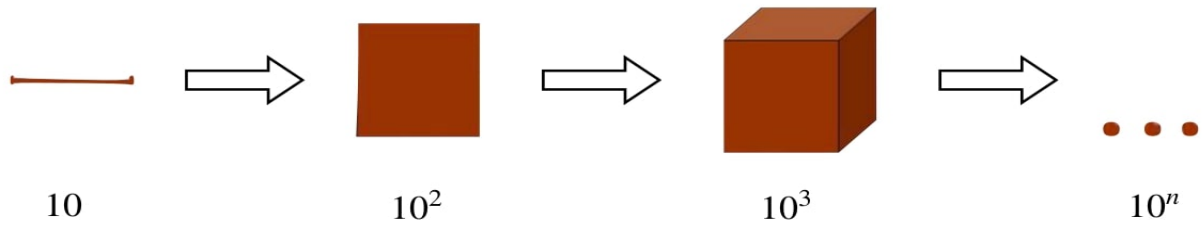


Рис. 1: Пример проклятия размерности.

Для устранения этой проблемы можно выбросить часть признаков или построить меньшее количество признаков на основе старых.

Цель: построить меньшее количество новых признаков, которые содержат максимум информации из исходных. Для начала рассмотрим Principal Component Analysis (PCA), метод главных компонент. Пусть есть матрица  $F_{l,n}$  где  $l$  - количество объектов,  $n$  - количество признаков. Хотим сократить количество признаков с  $n$  до  $m$ . При этом старые признаки должны как можно точнее линейно восстанавливаться по новым на обучающей выборке.

$$f_j^*(x) = \sum_{s=1}^m g_s(x) u_{js}, j = 1, \dots, n, \forall x \in X$$

Старые признаки  $f_j$  - линейная комбинация новых  $g_s$ .

Хотим восстанавливать старые признаки из новых, как можно точнее:

$$\sum_{i=1}^l \sum_{j=1}^n (f_j^*(x_i) - f_j(x_i))^2 \rightarrow \min_{\{g_s(x_i)\}, \{u_{js}\}}$$

Матрицы объекты признаки:

$$F_{l \times n} = \begin{bmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \vdots & \vdots & \vdots \\ f_1(x_l) & \dots & f_n(x_l) \end{bmatrix}$$

$$G_{l \times m} = \begin{bmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \vdots & \vdots & \vdots \\ g_1(x_l) & \dots & g_m(x_l) \end{bmatrix}$$

Матрица линейного преобразования новых признаков в старые:

$$U_{n \times m} = \begin{bmatrix} u_{11} & \dots & u_{1m} \\ \vdots & \vdots & \vdots \\ u_{n1} & \dots & u_{nm} \end{bmatrix}$$

$$F^* = GU^T$$

Мы хотим чтобы:

$$F^* = F$$

Из этого имеем:

$$\|GU^T - F\|^2 \rightarrow \min_{G,U}$$

**Теорема 1** Если  $m \leq rkF$ , то минимум  $\|GU^T - F\|^2$  достигается, когда столбцы  $U$  - это с.в. матрицы  $F^T F$ , соответствующие  $m$  максимальным с.з.  $\lambda_1, \dots, \lambda_m$ , а матрица  $G = FU$ . При этом:

1. матрица  $U$  ортонормирована:  $U^T U = I_m$ ;
2. матрица  $G$  ортогональна:  $G^T G = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ ;
3.  $U\Lambda = F^T F U$ ;  $G\Lambda = F F^T G$ ;
4.  $\|GU^T - F\|^2 = \|F\|^2 - \text{tr}\Lambda = \sum_{j=1}^n \lambda_j - \sum_{j=1}^m \lambda_j = \sum_{j=m+1}^n \lambda_j$ ;

Матрица  $G$  ортогональна  $G^T G = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$  а значит, новые признаки образуют базис.

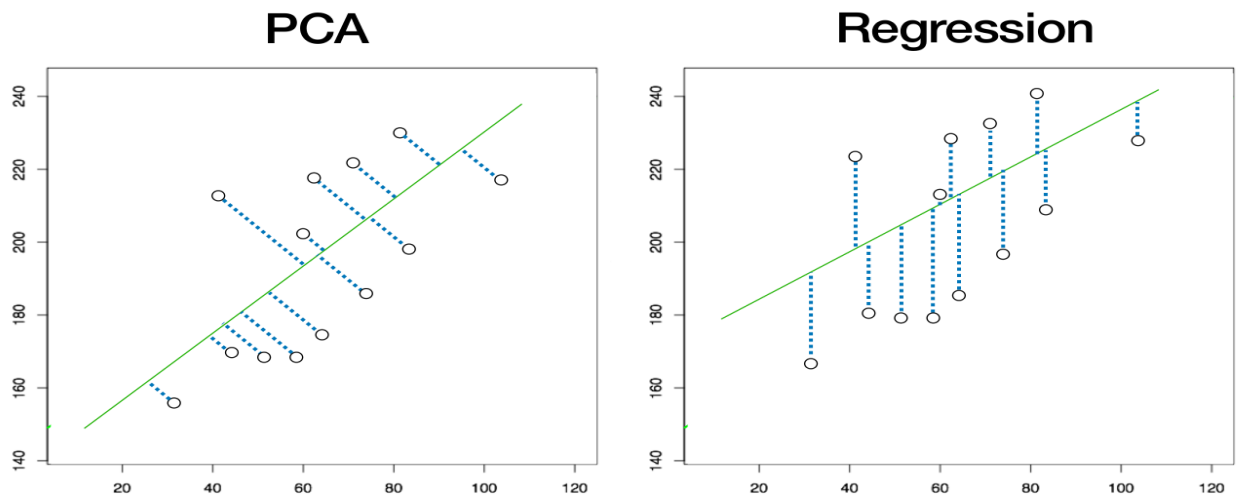


Рис. 2: Пример работы линейной регрессии и PCA.

Обобщая, PCA ищет такую гиперплоскость, суммарное расстояние от точек выборки до которых будет минимально. PCA ищет такие ортогональные проекции, дисперсия вдоль которых для точек выборки будет максимальна. PCA строит такой базис, в котором новые признаки ортогональны.

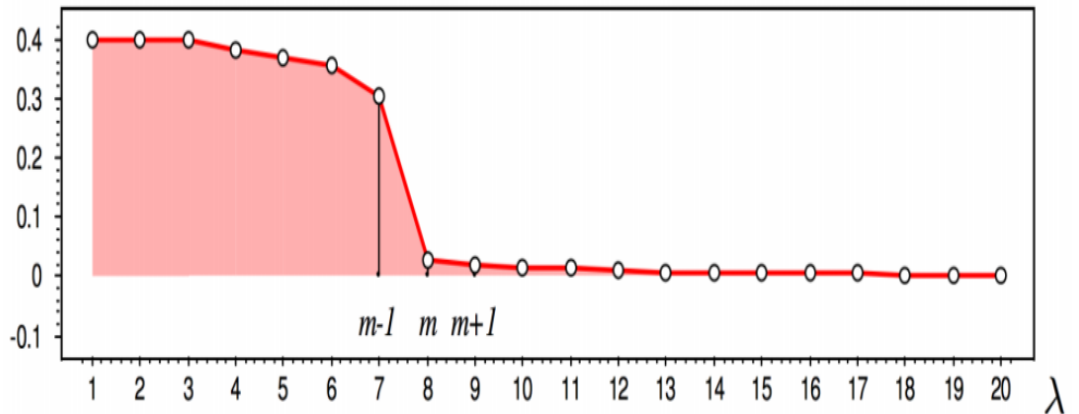
Допустим что мы построили такую плоскость. Как выбрать количество новых признаков? Собственные числа упорядочены по убыванию:

$$\lambda_1 \geq \dots \geq \lambda_n \geq 0.$$

$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \leq \varepsilon$$

$$\begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \lambda_m & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_{n-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & \lambda_n \end{pmatrix}$$

Критерий «крутого склона»: находим  $m$ :  $E_{m-1} \gg E_m$ :



Здесь видно, что есть некое значение  $m$ , такое что  $E_m$  намного меньше чем  $E_{m-1}$  именно это значение и будет количеством признаков, которые должны остаться в нашей таблице.

Теперь рассмотрим метод SVD (Singular Value Decomposition). Он основан на утверждении, что:

$$F_{l,n} = V_{l,n} * \Sigma_{n,n} * U_{n,n}^T$$

, где  $V$  и  $U$  - ортогональные матрицы (состоят из левых и правых сингулярных векторов),  $\Sigma$  - на главной диагонали сингулярные числа.

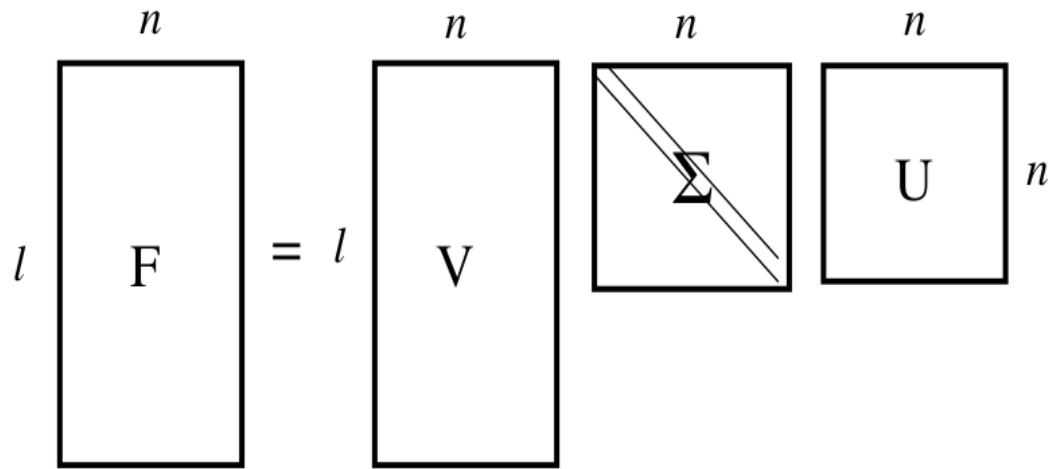
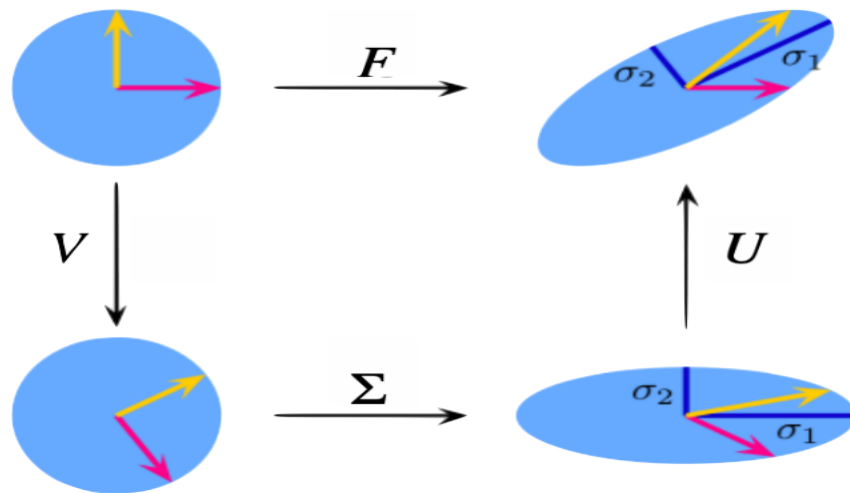


Рис. 3: Размеры матриц в SVD разложении.

Пусть исходная матрица  $F$  задает некоторое линейное преобразование исходного пространства. Тогда линейное преобразование можно разбить на последовательность более простых преобразований.

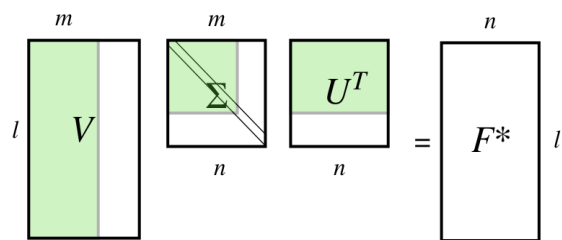


Можно взять  $m$  строк матрицы  $\Sigma$ , соответствующие наибольшим сингулярным числам,  $m$  столбцов матриц  $V$  и  $U \rightarrow$  получить приближенное представление матрицы  $F$ .

$$F_{l,n}^* = V_{l,m} * \Sigma_{m,m} * U_{m,n}^T$$

Сравним PCA и SVD.  
PCA:

$$F_{l,m} = G_{l,m} * U_{m,n}^T$$



SVD:

$$F_{l,n}^* = V_{l,m} * \Sigma_{m,m} * U_{m,n}^T$$

В качестве новых признаков можно взять произведение первых двух матриц:

$$G_{l,m} = V_{l,m} * \Sigma_{m,m}$$

