Метод главных компонент: постановка задачи

$f_1(x), \dots, f_n(x)$ — исходные числовые признаки;
$g_1(x), \dots, g_m(x)$ — новые числовые признаки, $m \leqslant n$;

Требование: старые признаки должны линейно восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, \quad j = 1, \dots, n, \quad \forall x \in X,$$

как можно точнее на обучающей выборке $x_1, \dots, x_\ell$:

$$\sum_{i=1}^{\ell} \sum_{j=1}^{n} \bigl(\hat{f}_j(x_i) - f_j(x_i)\bigr)^2 \to \min_{\{g_s(x_i)\}, \{u_{js}\}}$$

Матрицы «объекты–признаки», старая и новая:

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}; \quad \underset{\ell\times m}{G} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_\ell) & \dots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$\underset{n\times m}{U} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \qquad \hat{F} = GU^{\mathsf{T}} \overset{\text{хотим}}{\approx} F.$$

Найти: и новые признаки G, и преобразование U:

$$\sum_{i=1}^{\ell}\sum_{j=1}^{n}\bigl(\hat{f}_j(x_i) - f_j(x_i)\bigr)^2 = \bigl\| GU^{\mathsf{T}} - F \bigr\|^2 \to \min_{G,U},$$

Теорема

Если $m \leqslant \operatorname{rk} F$, *то минимум* $\|GU^{\mathsf{T}} - F\|^2$ *достигается, когда столбцы* U *— это с.в. матрицы* $F^{\mathsf{T}}F$, *соответствующие* m *максимальным с.з.* $\lambda_1, \ldots, \lambda_m$, *а матрица* $G = FU$.

При этом:

1. *матрица* U *ортонормирована:* $U^{\mathsf{T}}U = I_m$;
2. *матрица* G *ортогональна:* $G^{\mathsf{T}}G = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$;
3. $U\Lambda = F^{\mathsf{T}}FU$; $G\Lambda = FF^{\mathsf{T}}G$;
4. $\|GU^{\mathsf{T}} - F\|^2 = \|F\|^2 - \operatorname{tr}\Lambda = \sum\limits_{j=m+1}^{n} \lambda_j$.

Связь с сингулярным разложением

Если взять $m = n$, то:

1. $\left\|GU^\intercal - F\right\|^2 = 0$;
2. представление $\hat{F} = GU^\intercal = F$ точное и совпадает с сингулярным разложением при $G = V\sqrt{\Lambda}$:
$$F = GU^\intercal = V\sqrt{\Lambda}U^\intercal; \quad U^\intercal U = I_m; \quad V^\intercal V = I_m.$$
3. линейное преобразование U работает в обе стороны:
$$F = GU^\intercal; \quad G = FU.$$

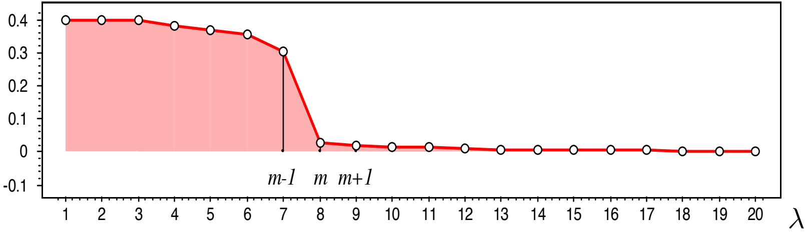
Поскольку новые признаки некоррелированы ($G^\intercal G = \Lambda$), преобразование U называется *декоррелирующим* (или преобразованием Карунена–Лоэва).

Упорядочим с.з. $F^{\mathsf{T}}F$ по убыванию: $\lambda_1 \geqslant \ldots \geqslant \lambda_n \geqslant 0$.

Эффективная размерность выборки — это наименьшее целое m, при котором

$$E_m = \frac{\|GU^{\mathsf{T}} - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\lambda_1 + \cdots + \lambda_n} \leqslant \varepsilon.$$

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



Задача наименьших квадратов для МЛР: $\|F\alpha - y\|^2 \to \min\limits_{\alpha}$.

Заменим $\underset{\ell\cdot n}{F}$ на её приближение $\underset{\ell\cdot m}{G} \cdot \underset{m\cdot n}{U^{\mathsf{T}}}$, предполагая $m \leqslant n$:

$$\|G\underbrace{U^{\mathsf{T}}\alpha}_{\beta} - y\|^2 = \|G\beta - y\|^2 \to \min_{\beta}.$$

Связь нового и старого вектора коэффициентов:

$$\beta = U^{\mathsf{T}}\alpha; \qquad \alpha = U\beta.$$

Решение задачи наименьших квадратов относительно β (единственное отличие — m слагаемых вместо n):

$$\beta^* = D^{-1}V^{\mathsf{T}}y; \qquad \alpha^* = UD^{-1}V^{\mathsf{T}}y = \sum_{j=1}^{m} \frac{1}{\sqrt{\lambda_j}} u_j(v_j^{\mathsf{T}}y);$$

$$G\beta^* = VV^{\mathsf{T}}y = \sum_{j=1}^{m} v_j(v_j^{\mathsf{T}}y);$$

- Метод главных компонент позволяет приближать матрицу её низкоранговым разложением.
- Для этого достаточно взять из SVD-разложения первые m сингулярных чисел и векторов матрицы.
- Этот приём широко используется в анализе данных — в задачах регрессии, классификации, сжатия данных, обработки изображений.