Linear Regression - Correlated features
One geometrical approach

Réda Arab

1 Empirical correlation and geometry

1.1 Link between scalar product and correlation

Suppose that we have two vectors U and V in $\mathbb{R}^n, n \geq 1$

$$U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

such that $\bar{U} = \frac{1}{n} \sum\limits_{i=1}^{n} u_i = 0\,,\, \bar{V} = \frac{1}{n} \sum\limits_{i=1}^{n} v_i = 0$

Therefore, the empirical covariance and correlation are :

$$cov(U, V) = \sum_{i=1}^{n} u_i v_i$$

$$corr(U, V) = \frac{\sum\limits_{i=1}^{n} u_i v_i}{\sqrt{\sum\limits_{i=1}^{n} u_i^2 \cdot \sum\limits_{i=1}^{n} v_i^2}} = r$$

Defining r as the empirical correlation, we can rewrite it as :

$$\boxed{r = \frac{\langle U, V \rangle}{\|U\|_2 \|V\|_2}} = \langle \frac{U}{\|U\|_2}, \frac{V}{\|V\|_2} \rangle$$

Therefore, if the empirical correlation r is near 1 (or -1), it means that U and V are closely aligned. Indeed, the scalar product of two unit vectors depends only on the angle between these two vectors (here $\frac{U}{\|U\|_2}$ and $\frac{V}{\|V\|_2}$).

1.2 Mathematical details

1.2.1 Projection on a vector

Let us define $P_V(U)$ the orthogonal projection of a vector U onto a vector V. We have $P_V(U) = \alpha \cdot V$, where α is a real number.

We can decompose U as $U = P_V(U) + \varepsilon$, where ε and V are orthogonal. Therefore :

$$\langle U, V \rangle = \langle P_V(U), V \rangle = \langle \alpha \cdot V, V \rangle = \alpha \|V\|_2^2$$

$$\text{so} \quad \alpha = \frac{\langle U, V \rangle}{\|V\|_2^2} \quad \text{and} \quad \boxed{P_V(U) = \frac{\langle U, V \rangle}{\|V\|_2^2} \cdot V}$$

We get :

$$U = P_V(U) + \varepsilon = \frac{\langle U, V \rangle}{\|V\|_2^2} V + \varepsilon \quad \text{with} \quad \langle V, \varepsilon \rangle = 0$$

i.e.

$$\boxed{U = r \cdot \frac{\|U\|_2}{\|V\|_2} V + \varepsilon}$$

After simplification, we obtain :

$$\|\varepsilon\|_2^2 = \|U - r \cdot \frac{\|U\|_2}{\|V\|_2} V\|_2^2$$

$$\boxed{\|\varepsilon\|_2^2 = (1 - r^2)\|U\|_2^2}$$

So if r is near 1 or -1, ε will have a norm near 0 and so U and V will be closely aligned.

1.2.2 Projection on a space generated by a set of vectors

Consider U a vector in $\mathbb{R}^n$ and $\mathcal{V} = \mathrm{span}(V_1, \ldots, V_k)$ where V_i in $\mathbb{R}^n$ for $i = 1, \ldots, k$ (all having mean 0).

Suppose there is $i \in \{1, \ldots, k\}$ such that :

$$\boxed{r_i = \frac{\langle U, V_i \rangle}{\|U\|_2 \|V_i\|_2} \approx 1 \,(\text{or} - 1)}$$

Therefore, we have shown in the previous section that :

$$U = P_{V_i}(U) + \varepsilon_i \quad \text{and} \quad \|\varepsilon_i\|_2^2 = (1 - r_i^2)\|U\|_2^2$$

where $P_{V_i}(U)$ is the orthogonal projection of U on V_i.

Let us denote $\Pi_{\mathcal{V}}(U)$ the orthogonal projection of U on $\mathcal{V}$, and we decompose U as $U = \Pi_{\mathcal{V}}(U) + \varepsilon$.

We know that $\Pi_{\mathcal{V}}(U)$ can be characterized as :

$$\Pi_{\mathcal{V}}(U) = \arg\min_{v \in \mathcal{V}} \|U - v\|_2^2$$

Therefore, $\|U - \Pi_{\mathcal{V}}(U)\|_2^2 \leq \|U - v\|_2^2$ for any v in $\mathcal{V}$.

In particular :

$$\|U - \Pi_{\mathcal{V}}(U)\|_2^2 = \|\varepsilon\|_2^2 \leq \|U - P_{V_i}(U)\|_2^2 = \|\varepsilon_i\|_2^2$$

So

$$\boxed{\|\varepsilon\|_2^2 \leq (1 - r_i^2)\|U\|_2^2}$$

If we have U and V_i closely aligned, we have U and $\mathcal{V}$ closely aligned (obvious geometrically speaking) and ε very small.

So if U and V_i are highly correlated, the projection of U on the orthogonal space of $\mathcal{V}$ (which is ε) will have a very low norm.

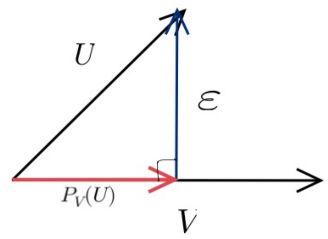

FIGURE 1 – Projection of U on a vector V

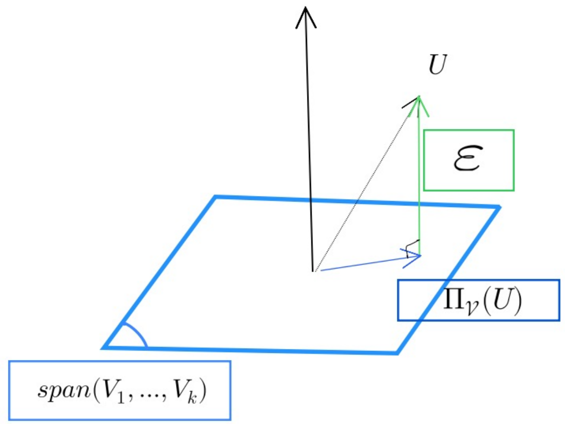

FIGURE 2 – Projection of U onto $span(V_1, \ldots, V_k)$

2 Recap : Ordinary Least Squares

2.1 Context and problem

Consider observations $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p, i = 1, \ldots, n$, and the aim is to infer a simple regression function relating the average value of a response, Y_i, and a collection of predictors or variables, x_i (i.e. regression task).

A linear model for the data assumes that it is generated according to

$$Y = X\beta^0 + \varepsilon$$

where $Y \in \mathbb{R}^n$ is the vector of responses ; $X \in \mathbb{R}^{n \times p}$ is the predictor matrix (or design matrix) with i th row $x_i^T; \varepsilon \in \mathbb{R}^n$ represents random error ; and $\beta^0 \in \mathbb{R}^p$ is the unknown vector of coefficients.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & \ldots & X_p \end{pmatrix}$$

Caution : ε here is the random error and should not be confused with the epsilon defined in the previous part which was in the decomposition of a vector into 2 orthogonal vectors. Here :

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Provided $p \leq n$ and X full column rank, we can estimate β by ordinary least squares (OLS). This leads to an estimator $\hat{\beta}^{\text{OLS}}$ with

$$\boxed{\hat{\beta}^{\text{OLS}} := \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2 = \left(X^T X\right)^{-1} X^T Y}$$

Under the assumptions that $\mathbb{E}\left(\varepsilon_i\right) = 0$ and $\operatorname{Var}(\varepsilon) = \sigma^2 I$ (and fixed design), we have :

- $\mathbb{E}_{\beta^0, \sigma^2}\left(\hat{\beta}^{\text{OLS}}\right) = \mathbb{E}\left\{\left(X^T X\right)^{-1} X^T \left(X\beta^0 + \varepsilon\right)\right\} = \beta^0.$
- $\operatorname{Var}_{\beta^0, \sigma^2}\left(\hat{\beta}^{\text{OLS}}\right) = \left(X^T X\right)^{-1} X^T \operatorname{Var}(\varepsilon) X \left(X^T X\right)^{-1} = \sigma^2 \left(X^T X\right)^{-1}.$

2.2 OLS and orthogonal projections

The fitted values, $\hat{Y} := X\hat{\beta}$ are then given by $X\left(X^T X\right)^{-1} X^T Y$.

We define P as $P := X\left(X^T X\right)^{-1} X^T$. It is an orthogonal projection onto the column space of X (P known as the 'hat' matrix because it puts the hat on Y).

Indeed, $P^T = P$, $P \circ P = P$ and $Im(P) = Im(X) = span(X_1, ..., X_p)$.

$(I - P)$, where I is the idendity matrix, is the orthogonal projection onto $Im(X)^\perp$, the orthogonal space of $Im(X)$.

N.B. An important point for the next step is the following : we often scale our columns before doing OLS (for example to use Gradient Descent more efficiently). So in general, the columns of X have mean 0 and we are in the context of the part 1. for which we suppose that the vectors have mean 0.

3 Another way of computing the estimates of the coefficients for OLS

Let us write X_j for the j^{th} column of X, and X_{-j} for the $n \times (p-1)$ matrix formed by removing the j^{th} column from X. Define P_{-j} as the orthogonal projection on to the column space of X_{-j} (i.e. the space generated by the $p-1$ other columns).

Proposition : Let $X_j^{\perp} := (I - P_{-j}) X_j$, so $X_j^{\perp}$ is *the orthogonal projection of X_j on to the orthogonal complement of the column space of X_{-j}*. Then

$$\hat{\beta}_j = \frac{(X_j^{\perp})^T Y}{\|X_j^{\perp}\|^2}$$

We have $\boxed{\operatorname{Var}\left(\hat{\beta}_j\right) = \sigma^2 \|X_j^{\perp}\|^{-2}}$.

Thus if X_j is closely aligned to the column space of X_{-j}, the variance of $\hat{\beta}_j$ will be large. In particular, if X_j and another columns X_i are highly correlated, the quantity $\|X_j^{\perp}\|^{-2}$ will be large and the variance also.

Indeed, we showed this in part 1.2.2 taking $U = X_j$, $\mathcal{V} = \operatorname{span}(X_1, \ldots, X_{j-1}, X_j, , X_p)$ and $\Pi_{\mathcal{V}}(U) = P_{-j}X_j$. So $\varepsilon = X_j^{\perp}$.

Proof. Note that $Y = PY + (I-P)Y$ and

$$X_j^T (I - P_{-j})(I-P)Y = X_j^T (I-P)Y = 0,$$

so

$$\frac{(X_j^{\perp})^T Y}{\|X_j^{\perp}\|^2} = \frac{(X_j^{\perp})^T X (X^T X)^{-1} X^T Y}{\|X_j^{\perp}\|^2}$$

Since $X_j^{\perp}$ is orthogonal to the column space of X_{-j}, we have

$$(X_j^{\perp})^T X = \left(0 \cdots 0 \, (X_j^{\perp})^T X_j 0 \cdots 0\right)$$

and $(X_j^{\perp})^T X_j = X_j^T (I - P_{-j}) X_j = \|(I - P_{-j}) X_j\|^2$

Conclusion : If a pair of variables are highly correlated with each other, the variances of the estimates of the corresponding coefficients will be large which is something that we want to avoid.

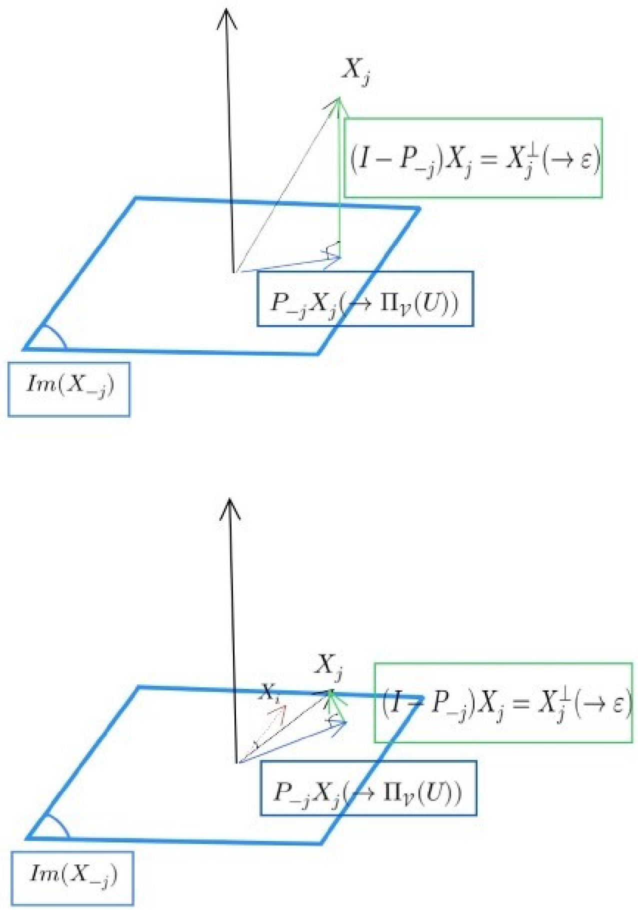

FIGURE 3 – Geometrical interpretation. X_i, a variable highly correlated with X_j, is added in the second scheme.

4 Appendix

Another way of seeing it, taking into account eigenvalues and SVD : `https://towardsdatascience.com/why-exclude-highly-correlated-features-when-building-regression-model-34d77a90ea8e`