Stats 202 PCA note

Jelena Markovic

October 23, 2018

Assume our data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ consists of n observations and p predictors. The elements of this matrix are denoted as x_{ij} – the element of i-th row and j-th column. Also, denote X_j, $j = 1, \ldots, p$, to be j-th column (predictor, feature) of matrix $\boldsymbol{X}$. For simplicity let us assume $p = 2$ so data consists of only two features and also that the features are centered so that $\frac{1}{n}\sum_{i=1}^{n} x_{i1} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} x_{i2} = 0$.

We want to see how the first principal component (PC1) looks like for $\boldsymbol{X}$. The first principal component is a two dimensional vector $\phi_1 = (\phi_{11}, \phi_{21})^\top$ given by solving

$$\max_{\phi_{11},\phi_{21}} \frac{1}{n}\sum_{i=1}^{n}(\phi_{11}x_{i1} + \phi_{21}x_{i2})^2 \qquad \text{s.t.} \qquad \phi_{11}^2 + \phi_{21}^2 = 0. \tag{1}$$

Let us discuss two ways to see the problem above:

1. Recall that a projection of a vector $a = (a_1, a_2)^\top$ onto a vector $b = (b_1, b_2)^\top$ is given by $\text{proj}_b a = \frac{a \cdot b}{b \cdot b} b = \frac{a_1 b_1 + a_2 b_2}{b_1^2 + b_2^2}(b_1, b_2)^\top$. Using this let us write a projection of the i-th data observation $x_i = (x_{i1}, x_{i2})^\top$ onto PC1 vector ϕ_1:

$$\text{proj}_{\phi_1} x_i = \frac{x_{i1}\phi_{11} + x_{i2}\phi_{21}}{\phi_{11}^2 + \phi_{21}^2}\phi_1 = (x_{i1}\phi_{11} + x_{i2}\phi_{21})\phi_1. \tag{2}$$

   The squared magnitude of this projection vector is $\|\text{proj}_{\phi_1} x_i\|_2^2 = (x_{i1}\phi_{11} + x_{i2}\phi_{21})^2$, which is one of the terms in the sum in the objective in (1). So we can write the PC1 objective from (1) as

$$\max_{\phi_{11},\phi_{21}} \frac{1}{n}\sum_{i=1}^{n} \|\text{proj}_{\phi_1} x_i\|_2^2 \qquad \text{s.t.} \qquad \phi_{11}^2 + \phi_{21}^2 = 0. \tag{3}$$

   Furthermore, denote with $r_i = (r_{i1}, r_{i2})^\top = x_i - \text{proj}_{\phi_1} x_i$ the residual after projecting vector x_i onto ϕ_1. Since $\|x_i\|_2^2 = x_{i1}^2 + x_{i2}^2 = \|\text{proj}_{\phi_1} x_i\|_2^2 + \|r_i\|^2$, we can write the above maximization problem as

$$\max_{\phi_{11},\phi_{21}} \frac{1}{n}\sum_{i=1}^{n}(\|x_i\|_2^2 - \|x_i - \text{proj}_{\phi_1} x_i\|_2^2) \qquad \text{s.t.} \qquad \phi_{11}^2 + \phi_{21}^2 = 0. \tag{4}$$

This problem is now equivalent to solving

$$\min_{\phi_{11},\phi_{21}} \frac{1}{n}\sum_{i=1}^{n} \|x_i - \text{proj}_{\phi_1} x_i\|_2^2 \qquad \text{s.t.} \qquad \phi_{11}^2 + \phi_{21}^2 = 0, \tag{5}$$

proving PC1 is the two-dimensional vector ϕ_1 minimizing the sum of squared distances from data points to the PC1 vector. This proves PC1 can be seen as in Figure 6.15 left in the book.

2. Another way to see (1) is by noticing that $\text{VAR}(\phi_{11}X_1+\phi_{21}X_2) = \mathbb{E}\left[(\phi_{11}X_1 + \phi_{21}X_2)^2\right] = \frac{1}{n}\sum_{i=1}^{n}(\phi_{11}x_{i1} + \phi_{21}x_{i2})^2$, where the variance and expectation are here with respect to the empirical (observed) distribution of the predictors X_1 and X_2. The PC1 objective becomes

$$\max_{\phi_{11},\phi_{21}} \text{VAR}(\phi_{11}X_1 + \phi_{21}X_2)^2 \qquad \text{s.t.} \qquad \phi_{11}^2 + \phi_{21}^2 = 0, \tag{6}$$

proving PC1 is a vector along which direction the variance of the features is maximized.