

# Principal Component Analysis

Sariel Har-Peled

10/12/2000

The following is taken from [Jol86].

## 1 Preliminaries

### 1.1 Probability

The *expectation* of a random variable  $X$  is:

$$E[X] = \sum_i p_i a_i,$$

where  $a_i$  are the possible values of  $X$ , and  $p_i$  is the probability of  $X$  to have the value  $a_i$ . Expectation is linear transformation; that is  $E[aX + b] = b + aE[x]$ . The *variance* of  $X$  is:

$$\text{var}(X) = E \left[ (X - E[X])^2 \right].$$

The *covariance* of two variables  $X, Y$  is

$$\text{cov}(X, Y) = E \left[ (X - E[X]) (Y - E[Y]) \right] = E \left[ XY \right] - E[X]E[Y] = \text{cov}(Y, X).$$

In particular,  $\text{cov}(X, X) = \text{var}(X)$ ,  $\text{cov}(aX + bY, Z) = a\text{cov}(X, Z) + b\text{cov}(Y, Z)$ . In particular, for random variables  $X_1, \dots, X_n$  we have:

$$\text{var} \left[ \sum_{i=1}^n a_i X_i \right] = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{cov}(X_i, X_j).$$

In particular, let  $S$  be the  $n \times n$  matrix, where the  $i, j$  entry, is  $s_{ij} = \text{cov}(X_i, X_j)$ . The matrix  $S$  is the *covariance* matrix, and in particular, if  $\mathbf{x} = (a_1 X_1, \dots, a_n X_n)$ , then

$$\text{var} \left[ \sum_{i=1}^n a_i X_i \right] = \mathbf{x}^T S \mathbf{x}.$$

## 1.2 Lagrange Multipliers

The following can be found in standard books about calculus. Let  $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$  be two given functions, and we wish to solve the following optimization problem:

$$\max_{\mathbf{x}} f(\mathbf{x}) \quad (1)$$

$$\text{s.t. } g(\mathbf{x}) = c, \quad (2)$$

where  $c$  is a prescribed constant. Observe, that the set of points that comply with the constraint  $g(\mathbf{x}) = c$  is a  $d - 1$ -dimensional surface  $\mathcal{C}$  in  $\mathbb{R}^d$ . In particular, let  $\mathbf{x}_0$  be the point that realizes the above maximum, and let  $\gamma(t) = (\gamma^1(t), \dots, \gamma^n(t))$  be a curve on the surface  $\mathcal{C}$  that passes through the point  $\mathbf{x}_0 = \gamma(t_0)$ . In particular, the function  $g(\gamma(t)) = c$  for every value of  $t$ . Namely,  $(g(\gamma(t)))' = 0$ . Using the chain rule, we know:

$$(g(\gamma(t)))' = \sum_{i=1}^n \frac{\partial g(\gamma(t))}{\partial x_i} \frac{dx_i}{dt} = 0,$$

where  $\frac{\partial g(\mathbf{x})}{\partial x_i}$  denote the  $i$ -th variable derivative of  $g$ , and  $\frac{dx_i}{dt} = \gamma'_i(t)$ . In particular, let

$$\nabla g(\mathbf{x}) = \left( \frac{\partial g(\mathbf{x})}{\partial x_1}, \frac{\partial g(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial g(\mathbf{x})}{\partial x_n} \right)$$

denote the *gradient* vector of  $g$  at the  $\mathbf{x}$ , and let  $\gamma'(t) = (\gamma'_1(t), \gamma'_2(t), \dots, \gamma'_n(t))$ . In particular,

$$\nabla g(\gamma(t_0)) \cdot \gamma'(t_0) = 0.$$

Note, that  $\gamma$  was a completely arbitrary curve along  $\mathcal{C}$  that passes through  $\mathbf{x}_0$ . We thus conclude, that  $\nabla g(\mathbf{x}_0)$  is the normal to  $\mathcal{C}$  at  $\mathbf{x}_0$ .

On the other side, we know that  $f(\mathbf{x})$  is being maximized at  $\mathbf{x}_0$ . In particular,  $f(\gamma(t))$  is maximized at  $t_0$ , which implies that  $(f(\gamma(t_0)))' = 0$ . Arguing, as above, we conclude that

$$\nabla f(\gamma(t_0)) \cdot \gamma'(t_0) = 0.$$

Which implies that

$$\nabla f(\gamma(t_0)) \cdot \gamma'(t_0) = \nabla g(\gamma(t_0)) \cdot \gamma'(t_0).$$

Namely, as  $\gamma$  was arbitrary, it implies that  $\nabla f(\mathbf{x}_0)$  is parallel to  $\nabla g(\mathbf{x}_0)$ ; namely, there exists a constant  $\lambda$  (called the *Lagrange multiplier*) such that  $\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0)$ . Thus, to solve the optimization problem Equation (2), it is enough to solve the following system of equations:

$$\text{for } i = 1, \dots, n \quad \frac{\partial f(\mathbf{x})}{\partial x_i} = \lambda \frac{\partial g(\mathbf{x})}{\partial x_i} \quad (3)$$

$$g(\mathbf{x}) = c \quad (4)$$

Note, that the solution to Equation (4) involves one additional variable (i.e.,  $\lambda$ ), and the solutions to Equation (4) are a superset of the solutions to Equation (2).

## 2 PCA - Principal Component Analysis

Given a covariance matrix  $S$  of  $n$  random variables  $X_1, \dots, X_p$ , find the combination of variables  $Y = \sum_{i=1}^p a_i X_i$  that *maximizes* variance. Namely,  $\text{var}(Y)$  is maximum. Of course, we need to normalize the coefficients  $a_i$ . In particular, let  $\mathbf{a} = (a_1, \dots, a_p)$ , we require that  $\|\mathbf{a}\| = 1$ , or alternatively,  $\mathbf{a}^T \mathbf{a} = 1$ . So, we have to solve the following optimization problem:

$$\begin{aligned} f(\mathbf{a}) = \max_{\mathbf{a}} \text{var}(\mathbf{a}^T \mathbf{x}) &= \max_{\mathbf{a}} \mathbf{a}^T S \mathbf{a} = \sum_{i,j} a_i a_j \text{cov}(X_i, X_j) \\ \text{s.t. } g(\mathbf{a}) = \mathbf{a}^T \mathbf{a} &= \sum_i a_i^2 = 1, \end{aligned}$$

where  $\mathbf{x} = (X_1, \dots, X_p)$ . By the Lagrange multipliers technique, this can be solved by solving:

$$\begin{aligned} \text{for } i = 1, \dots, p \quad \frac{\partial f(\mathbf{a})}{\partial a_i} &= \lambda \frac{\partial g(\mathbf{a})}{\partial a_i} \\ \text{s.t. } g(\mathbf{a}) &= 1 \end{aligned}$$

or equivalently,

$$\begin{aligned} \text{for } i = 1, \dots, p \quad \frac{\partial \sum_{i,j} a_i a_j \text{cov}(X_i, X_j)}{\partial a_i} &= \lambda \frac{\partial \sum_i a_i^2}{\partial a_i} \\ \text{s.t. } \sum_i a_i^2 &= 1. \end{aligned}$$

Which is

$$\text{for } i = 1, \dots, p \quad 2 \sum_{j=1}^p a_j \text{cov}(X_i, X_j) = \lambda 2a_i \quad \text{s.t.} \quad \sum_i a_i^2 = 1.$$

Alternatively,

$$\text{for } i = 1, \dots, p \quad \vec{S}_i \mathbf{a} = \sum_{j=1}^n a_j \text{cov}(X_i, X_j) - \lambda a_i = 0 \quad \text{s.t.} \quad \sum_i a_i^2 = 1.$$

where  $\vec{S}_i$  is the  $i$ -th line of the matrix  $S$ . Namely, we can rewrite this as,

$$(S - \lambda I) \mathbf{a} = 0 \quad \text{s.t.} \quad \sum_i a_i^2 = 1.$$

Namely,  $\lambda$  is an eigenvalue of  $S$ . But what eigenvalue? Observe, that the solution vector  $\alpha_1$  to the above system is an eigenvector, and in particular, the quantity we try to maximize is

$$\alpha_1^T S \alpha_1 = \lambda \alpha_1^T \alpha_1 = \lambda.$$

Thus, in choosing the  $\lambda$ , we should pick the largest eigenvalue.

Thus, by computing the largest eigenvalue, we had computed the combination of variables that maximizes the variation. Next, we want to find a combination of variables that is uncorrelated

with  $\mathbf{a}^T \mathbf{x}$  that maximizes the variation. Arguing as above, it must be an eigenvector, and it must be perpendicular to the previous eigenvector. Namely, the second combination of variables that maximizes the correlation is the second eigenvector of  $S$  corresponding to the second largest eigenvalue of  $S$ .

Let  $\alpha_1, \dots, \alpha_p$  denotes those (normalized!) eigenvectors. Those vectors are called the *principal components* (PC) of  $S$ . Furthermore,

$$\text{var}(\alpha_i^T \mathbf{x}) = \alpha_i^T S \alpha_i = \lambda_i,$$

which is the  $i$ -th largest eigenvalue of  $S$ .

## References

- [Jol86] I.T. Jolliffe. *Principal Component Analysis*. Springer series in statistics. Springer-Verlag, 1986. ISBN: 0387962697.