PCA Derivation

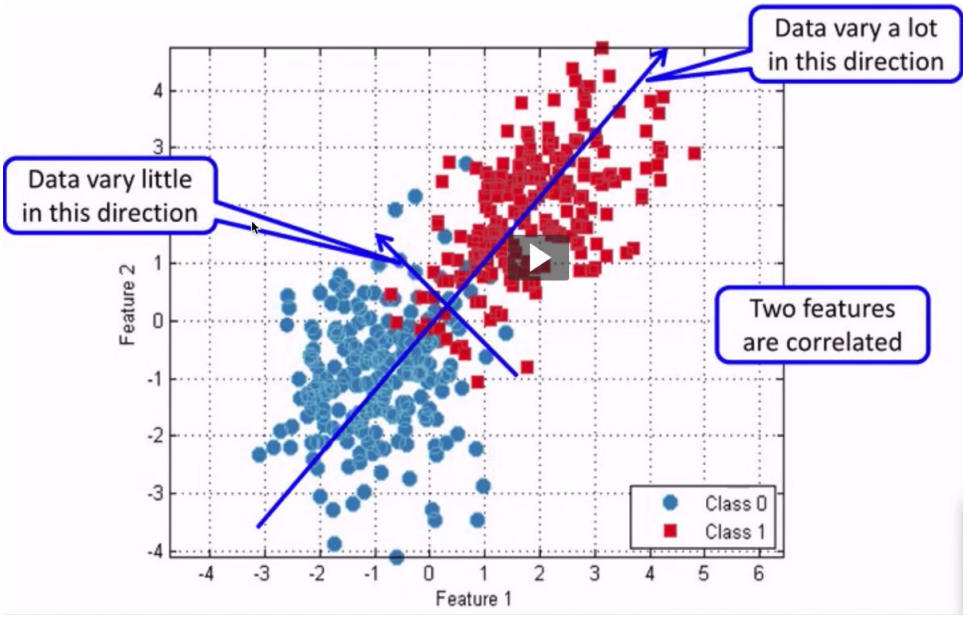
Garrett N. Bushnell

May 2021

1 Background

Suppose we're given data in a high-dimensional space and wish to convert it into lower dimensional. To preserve the most "signal", we want to ensure we capture the most variance.

For our example, suppose we have two features (Feature 1 and Feature 2), that are both in R^2. We wish to map our data into R^1, or a line. Which line should we choose? Per above, we want to choose the one with the most variance. So we'll choose the longest line.



Our goal is to map (aka, project, aka Inner Product) our data onto this chosen line, reducing it's dimension and keeping it's information. We note that this "line" is really just a linear combination of our features.

But how do we find this optimal combination of our features (w)? (Assuming we don't know that PCA is our solution)

2 Mathematical Formulation

So how do we find the optimal w? Since we're trying to maximize the variance, we'll need the mean(μ) and variance(C) of our data.

$$\begin{aligned} \mu &= \frac{1}{m}\sum_{i=1}^{m} x^i \\ C &= \frac{1}{m}\sum_{i=1}^{m} (x^i - \mu)(x^i - \mu)^T \end{aligned} \tag{2.1}$$

Our goal is to find the w that maximizes the variance (C). First, we need to re-write our μ and C in terms of w.

$$\begin{aligned} \mu &= \frac{1}{m}\sum_{i=1}^{m} wx^i \\ C &= \frac{1}{m}\sum_{i=1}^{m} (wx^i - w\mu)(wx^i - w\mu)^T \end{aligned} \tag{2.2}$$

$$\begin{aligned} &\max_w C \\ &\max_w \frac{1}{m}\sum_{i=1}^{m} (wx^i - w\mu)(wx^i - w\mu)^T \\ &\max_w \frac{1}{m}\sum_{i=1}^{m} (w^T x^i - w^T \mu)^2 \end{aligned} \tag{2.3}$$

So our goal is to maximize the above, but if we look, we realize that if we just keep increasing w, we'll also just keep increasing our objective function. We need to introduce a limit on w - that the norm is restricted to 1

$$\max_{w:\|w\|\leq 1} \frac{1}{m}\sum_{i=1}^{m} (w^T x^i - w^T \mu)^2 \tag{2.4}$$

We can now manipulate this equation to come up with a simpler solution

$$\begin{aligned} &= \max_{w:\|w\|\leq 1} \frac{1}{m}\sum_{i=1}^{m} (w^T x^i - w^T \mu)^2 \\ &= \max_{w:\|w\|\leq 1} \frac{1}{m}\sum_{i=1}^{m} (w^T (x^i - \mu))^2 \\ &= \max_{w:\|w\|\leq 1} \frac{1}{m}\sum_{i=1}^{m} (w^T (x^i - \mu) \cdot (x^i - \mu)^T w) \\ &= \max_{w:\|w\|\leq 1} w^T \left(\frac{1}{m}\sum_{i=1}^{m} ((x^i - \mu) \cdot (x^i - \mu)^T \right) w \\ &= \max_{w:\|w\|\leq 1} w^T (C)\, w \end{aligned} \tag{2.5}$$

Now that our equation is simpler (we've separated our objective (w) from the rest), we can focus on optimizing it. We'll need to use Lagrangian Multipliers.

3 Solution

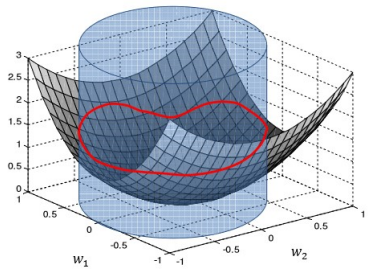
For simplicity, let's assume we're given an original dataset consisting of 2 feature (R^2), and that their covariance matrix (C) is given by

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \tag{3.1}$$

Our optimization problem becomes

$$\begin{aligned} &\max_{w:\|w\|\leq 1} \; w^T (C)\, w \\ &\max_{w:\|w\|\leq 1} \; \begin{bmatrix} w_1, w_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\ &\max_{w:\|w\|\leq 1} \; w_1^2 + 2w_2^2 \end{aligned} \tag{3.2}$$

We can visualize our problem (with the red circle being our constraint on w).



Since we now have our constrained opimization problem, we can use the Lagrangian function to optimize.

$$\begin{aligned} &\max_{w:\|w\|\leq 1} \; w^T (C)\, w \\ &L(w, \lambda) = w^T C w + \lambda(1 - \|w\|^2) \end{aligned} \tag{3.3}$$

We'll want to differentiate with respect to w and set it equal to zero to solve for the solution.

$$\begin{aligned} \frac{\partial L}{\partial w} &= 2Cw - 2\lambda w \\ 2Cw &= 2\lambda w \\ Cw &= \lambda w \end{aligned} \tag{3.4}$$

The above is the exact definition of an Eigendecomposition. Hence, the Principal Components are the solution to our problem. If we want to convert to a 1-D space, we only select the first Eigenvector. If we want to reduce to a 3-D space, we need to select the first 3 Eigenvectors.

We then need to project our data onto our Eigenvectors for the final solution - the dimensional reduced data matrix Z

$$z^i = \begin{pmatrix} w^{1\top}(x^i - \mu)/\sqrt{\lambda_1} \\ w^{2\top}(x^i - \mu)/\sqrt{\lambda_2} \\ \vdots \end{pmatrix}$$

A visual representation of our solution is below.