# Principal Components Analysis in 2D

Miguel A. Lerma

October 30, 2017

**Abstract**

Here we study 2-dimensional PCA and discuss an application to the location of a set of points in the plane in an elliptical region.

## 1 Introduction

Principal component analysis (PCA) is a mathematical procedure intended to replace a number of correlated variables with a new set of variables that are linearly uncorrelated. This can be achieved with a change of variables given by an orthogonal transformation. The technique is well known, however the case of two variables is particularly simple to analyze, and has some interesting application to image analysis. Here we analyze the math behind the method, which becomes particularly simple to approach in two dimensions, and use it to solve the problem of finding a elliptical region approximately encompassing a given set of points. In our example we will use the method to image (1) consisting of a rotated "HELLO" word.

## 2 The Math behind PCA

Assume that we have a collection of 2-dimensional points $(x_1, y_1)$, $(x_2, y_2)$,..., $(x_n, y_n)$. The correlation between the $x_i$ and the $y_i$ can be measured by their variance $\sigma_{xy} = \overline{(x-\bar{x})(y-\bar{y})}$, where $\overline{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$ represents the average value of $u$. In order to simplify computations we will assume $\bar{x} = \bar{y} = 0$, which can be accomplished with a translation $(x, y) \mapsto (x - \bar{x}, y - \bar{y})$.


Figure 1: A rotated "HELLO"

Our purpose is to find a change of coordinates $(x, y) \mapsto (x', y')$ given by a rotation:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = R \begin{pmatrix} x \\ y \end{pmatrix}, \qquad R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

such that $x'$ and $y'$ are uncorrelated, i.e., $\sigma_{x'y'} = 0$.

Let $X$ be the matrix whose columns are the original coordinates of the given points:

$$X = \begin{pmatrix} x_1 & x_2 & \cdots x_n \\ y_1 & y_2 & \cdots y_n \end{pmatrix}$$

and $X'$ is the analogous matrix with the new coordinates:

$$X' = \begin{pmatrix} x'_1 & x'_2 & \cdots x'_n \\ y'_1 & y'_2 & \cdots y'_n \end{pmatrix} = RX\,.$$

The correlation matrix of $X$ is

$$C = \frac{1}{n} X X^T = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix},$$

and that of $X'$ is

$$C' = \frac{1}{n} X' X'^T = \frac{1}{n} R X X^T R^T = R C R^T\,,$$

Since $R$ is a rotation it is orthogonal and we have $R^T = R^{-1}$. On the other hand we want $\sigma_{x'y'} = 0$, hence $C'$ must be a diagonal matrix:

$$C' = \begin{pmatrix} \sigma_{x'}^2 & 0 \\ 0 & \sigma_{y'}^2 \end{pmatrix}.$$

So the problem amounts to diagonalizing $C$. This can be done in the usual way, by finding the eigenvalues and eigenvectors of $C$.

The eigenvalues of $C$ are the roots of its characteristic polynomial:

$$\begin{vmatrix} \sigma_x^2 - \lambda & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 - \lambda \end{vmatrix} = \lambda^2 - (\sigma_x^2 + \sigma_y^2)\lambda + \sigma_x^2\sigma_y^2 - \sigma_{xy}^2$$

The roots can be found with the usual formula to solve second degree equations:

$$\begin{aligned} \lambda &= \frac{\sigma_x^2 + \sigma_y^2 \pm \sqrt{(\sigma_x^2 + \sigma_y^2)^2 - 4\sigma_x^2\sigma_y^2 + 4\sigma_{xy}^2}}{2} \\ &= \frac{\sigma_x^2 + \sigma_y^2 \pm \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}}{2} . \end{aligned}$$

Let $\lambda_+$ be the root obtained using the plus sign in the formula, and $\lambda_-$ the one obtained with the minus sign. Then it can be easily checked that the following are eigenvectors verifying $CV_\pm = \lambda_\pm V_\pm$:

$$V_+ = \begin{pmatrix} \sigma_x^2 + \sigma_{xy} - \lambda_- \\ \sigma_y^2 + \sigma_{xy} - \lambda_- \end{pmatrix} \qquad V_- = \begin{pmatrix} \sigma_x^2 + \sigma_{xy} - \lambda_+ \\ \sigma_y^2 + \sigma_{xy} - \lambda_+ \end{pmatrix}$$

These vectors can now be normalized $U_+ = k_+V_+$, $U_- = k_-V_-$, with $k_+$ and $k_+$ chosen so that $||U_+|| = ||U_-|| = 1$, and $U_+ \times U_- = 1$. If we choose the rotation matrix $R$ so that

$$R^T = (U_+|U_-) = \begin{pmatrix} U_{+,x} & U_{-,x} \\ U_{+,y} & U_{-,y} \end{pmatrix} ,$$

then $\lambda_+ = \sigma_{x'}^2$, and $\lambda_- = \sigma_{y'}^2$. Since $\lambda_+ \geq \lambda_-$ we will have that $x'$ will be the variable with the maximum dispersion, and $y'$ the one with minimum dispersion.

# 3 Application to optimal location of a set of points in an elliptical region

We now apply the method to locating the position, orientation, and approximate size of a set of points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ in the plane, and illustrate it for the example shown in figure (1). The set of points in

this example is the one given by the black points of the word "HELLO" in the image. Each has some coordinates $(x_i, y_i)$, and its barycenter $(\bar{x}, \bar{y})$ is the first reference we use for the location of the set on the plane. The eigenvectors $V_+$ and $V_-$ determine the directions of maximum and minimum dispersion, and their directions can be used for the axes of an ellipse centered in the barycenter. It remains the determine the values of the semi axes of the ellipse, which will be proportional to the standard deviations $\sigma_{x'} = \sqrt{\lambda_+}$ and $\sigma_{y'} = \sqrt{\lambda_-}$ respectively. In this particular example the sizes has been chosen to be double of the standard deviations, and the result can be seen in figure (2). The ellipse has been drawn in green, and the eigenvectors, scaled so that their lengths coincide with double the standard deviations of $x'$ and $y'$, are drawn in red and yellow respectively.


Figure 2: The rotated "HELLO" and its associated ellipse

See the appendix for specific code designed to perform the computations. The program used in the example shown here was actually implemented in Matlab, and has not been included here, but the computations are basically identical to the ones performed by the code included below.

## 4 Conclusion

We have reviewed the mathematical details of PCA in two dimensions, and used it to solve a problem in image analysis, namely locating a set of points in the plane in an elliptical region.

## Appendix: sample code implementation

The following is C computer code illustrating how to perform the computations (based on tested code but modified and simplified to be included here, it may contain typos or bugs):

```
// some variable declarations
long int sumx=0, sumy=0, sumxx=0, sumyy=0, sumxy=0;
double xbar, ybar, varx, vary, covxy, sumvars, diffvars;
double discriminant, sqrtdiscr, lambdaplus, lambdaminus;
double aplus, bplus, aminus, bminus;

// scan the image (nLength and nHeight are the dimensions of the image)
for(x=0; x<nLength; x++)
{
     for(y=0; y<nHeight; y++)
     {
          if (region(x,y)) // true is we are in the region of interest
          {
              sumx+=x; sumy+=y; sumxx+=x*x;
              sumyy+=y*y; sumxy+=x*y; npix++;
          }
     }
}

// baricenter
xbar = ((double)sumx)/npix;
ybar = ((double)sumy)/npix;

// variances and covariance
varx = sumxx/npix - xbar*xbar;
vary = sumyy/npix - ybar*ybar;
covxy = sumxy/npix - xbar*ybar;

sumvars = varx + vary;
diffvars = varx - vary;
discriminant = diffvars*diffvars + 4*covxy*covxy;
sqrtdiscr = sqrt(discriminant);
```

```
// eigenvalues
lambdaplus  = (sumvars + sqrtdiscr)/2;
lambdaminus = (sumvars - sqrtdiscr)/2;

//eigenvectors - these are the components of the two vectors
aplus = varx + covxy - lambdaminus;
bplus = vary + covxy - lambdaminus;

aminus =  varx + covxy - lambdaplus;
bminus =  vary + covxy - lambdaplus;

// normalizing the vectors
double aParallel;       double bParallel;
double aNormal;         double bNormal;

double denomPlus = sqrtf(aplus*aplus + bplus*bplus);
double denomMinus= sqrtf(aminus*aminus + bminus*bminus);

aParallel = aplus/denomPlus;
bParallel = bplus/denomPlus;
aNormal = aminus/denomMinus;
bNormal = bminus/denomMinus;

// semi axes
double k = 2 // scale factor
double majoraxis = k*sqrtf(lambdaplus);
double nimoraxis = k*sqrtf(lambdaminus);
```