CSE/ECE-471: Statistical Methods in AI **Spring 2020**

Principal Components Analysis

Prepared by: Yallamanda Rao (2019201029),Vikram Keswani(2019201059),keshav (20161150)

Principal components analysis(PCA) is one of a family of techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information.PCA is one of the simplest and most robust ways of doing such dimensionality reduction.

1 Background Mathematics

1.1 Variance

variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value.the formulae for computing variance is:

$$\text{Var(X)} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

1.2 Covariance

Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

The formula for covariance is very similar to the formula for variance. The formula for variance could also be written like this:

$$\text{Var(X)} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}{n-1}$$

The formulae for covariance is:

$$\text{Covar(X,Y)} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

1.3 Covariance Matrix

If we have a data set with more than 2 dimensions, there is more than one covariance measurement that can be calculated.culated. For example, from a 3 dimensional data set (dimensions x,y,z) you could calculate cov(x,y),cov(y,z),cov(x,z).

A useful way to get all the possible covariance values between all the different dimensions is to calculate them all and put them in a matrix.The definition for the covariance matrix for a set of data with n dimensions is:

$$C^{nxn} = (c_{i,j}, c_{i,j} = cov(Dim_x, Dim_y))$$

where C^{nxn} is a matrix with n rows and n columns and Dim_x is xth dimension.

For a 3 dimensional dataset with dimensions x,y,z the covariance matrix will be as below

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix}$$

Some points to note: Down the main diagonal, you see that the covariance value is between one of the dimensions and itself. These are the variances for that dimension.The other point is that since cov(a,b) = cov(b,a) the matrix is symmetrical about the main diagonal.

We can compute covariance matrix easily.Let our data be expressed as matrix X $\in R^{pxn}$.Initially we have to compute the mean vector.And then subtract mean vector from each data point in X.Note that performing this operation will make mean of vectors zero($\frac{1}{n}\sum_{i=1}^{n} x_i = 0$) .We can then obtain co variance matrix easily by doing $\frac{1}{n-1}\mathrm{X}^T X$. Let v be the co variance matrix.

2 Mathematics of PCA

We start with p-dimensional vectors, and want to summarize them by projecting down into a q-dimensional subspace. Our summary will be the projection of the original vectors on to q directions, the principal components, which span the subspace. There are several equivalent ways of deriving the principal components mathematically. The simplest one is by finding the projections which maximize the variance. The first principal component is the direction in space along which projections have the largest variance. The second principal component is the direction which maximizes variance among all directions orthogonal to the first.The kth component is the variance-maximizing direction orthogonal to the previous k 1 components.There are p principal components in all.

2.1 Minimizing Projection Residuals

We'll start by looking for a one-dimensional projection. That is, we have p-dimensional vectors, and we want to project them on to a line through the origin. We can specify the line by a unit vector along it,$\vec{w}$,and then the projection of a data vector $\vec{x_i}$ on to the line is $\vec{x_i}.\vec{w}$ which is a scalar. This is the distance of the projection from the origin; the actual coordinate in p-dimensional space is $(\vec{x_i}.\vec{w})\vec{w}$.The mean of the projections will be zero, because the mean of the vectors $\vec{x_i}$ is zero:

$$\frac{1}{n}\sum_{i=1}^{n}(\vec{x_i}\cdot\vec{w})\vec{w} = \left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\cdot\vec{w}\right)\vec{w}$$

If we try to use our projected or image vectors instead of our original vectors, there will be some error, because (in general) the images do not coincide with the original vectors. (When do they coincide?) The

$$\begin{aligned}
\|\vec{x_i} - (\vec{w} \cdot \vec{x_i})\vec{w}\|^2 &= (\vec{x_i} - (\vec{w} \cdot \vec{x_i})\vec{w}) \cdot (\vec{x_i} - (\vec{w} \cdot \vec{x_i})\vec{w}) \\
&= \vec{x_i} \cdot \vec{x_i} - \vec{x_i} \cdot (\vec{w} \cdot \vec{x_i})\vec{w} \\
&\quad -(\vec{w} \cdot \vec{x_i})\vec{w} \cdot \vec{x_i} + (\vec{w} \cdot \vec{x_i})\vec{w} \cdot (\vec{w} \cdot \vec{x_i})\vec{w} \\
&= \|\vec{x_i}\|^2 - 2(\vec{w} \cdot \vec{x_i})^2 + (\vec{w} \cdot \vec{x_i})^2 \vec{w} \cdot \vec{w} \\
&= \vec{x_i} \cdot \vec{x_i} - (\vec{w} \cdot \vec{x_i})^2
\end{aligned}$$

difference is the error or residual of the projection. How big is it? For any one vector, say $\vec{x_i}$ it's since w is unit vector. Adding these residuals for all vectors we get MSE

$$\begin{aligned}
MSE(\vec{w}) &= \frac{1}{n}\sum_{i=1}^{n} \|\vec{x_i}\|^2 - (\vec{w} \cdot \vec{x_i})^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} \|\vec{x_i}\|^2 - \sum_{i=1}^{n} (\vec{w} \cdot \vec{x_i})^2\right)
\end{aligned}$$

The first summation doesn't depend on $\vec{w}$, so it doesn't matter for trying to minimize the mean squared residual. To make the MSE small, what we must do is make the second sum big, i.e., we want to maximize

$$\frac{1}{n}\sum_{i=1}^{n} (\vec{w} \cdot \vec{x_i})^2$$

which we can see is the sample mean of $(\vec{x_i}.\vec{w})^2$. The mean of a square is always equal to the square of the mean plus the variance:

$$\frac{1}{n}\sum_{i=1}^{n} (\vec{w} \cdot \vec{x_i})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} \vec{x_i} \cdot \vec{w}\right)^2 + \mathrm{Var}\left[\vec{w} \cdot \vec{x_i}\right]$$

Since we've just seen that the mean of the projections is zero, minimizing the residual sum of squares turns out to be equivalent to maximizing the variance of the projections.

2.2 Maximizing Variance

Accordingly, let's maximize the variance! Writing out all the summations grows tedious, so let's do our algebra in matrix form. If we stack our n data vectors into an $n \times p$ matrix, x, then the projections are given by xw, which is an $n \times 1$ matrix. The variance is

$$\begin{aligned}
\sigma^2_{\vec{w}} &= \frac{1}{n}\sum_i \left(\vec{x_i} \cdot \vec{w}\right)^2 \\
&= \frac{1}{n}(\mathbf{xw})^T(\mathbf{xw}) \\
&= \frac{1}{n}\mathbf{w}^T\mathbf{x}^T\mathbf{xw} \\
&= \mathbf{w}^T\frac{\mathbf{x}^T\mathbf{x}}{n}\mathbf{w} \\
&= \mathbf{w}^T\mathbf{vw}
\end{aligned}$$

We want to chose a unit vector so as to maximize variance.To do this, we need to make sure that we only look at unit vectors — we need to constrain the maximization. The constraint is that $w^T w = 1$.we can do this by introducing a new variable, the Lagrange multiplier , adding times the constraint equation to our objective function, and doing an unconstrained optimization. For our projection problem,

$$\begin{aligned}
\mathscr{L}(\mathbf{w},\lambda) &\equiv \sigma^2_{\mathbf{w}} - \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
\frac{\partial L}{\partial \lambda} &= \mathbf{w}^T\mathbf{w} - 1 \\
\frac{\partial L}{\partial \mathbf{w}} &= 2\mathbf{vw} - 2\lambda\mathbf{w}
\end{aligned}$$

Setting the derivatives to zero at the optimum, we get

$$\begin{aligned}
\mathbf{w}^T\mathbf{w} &= 1 \\
\mathbf{vw} &= \lambda\mathbf{w}
\end{aligned}$$

Thus, desired vector w is an eigenvector of the covariance matrix v, and the maximizing vector will be the one associated with the largest eigenvalue λ.

3 STEPS TO PERFORM

Here we will go through the step to perform the PCA on the set

3.1 GET SOME DATA

I am going to use the own made up data set with 2 dimensions shown below

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Dat

Original PCA data

3.2 SUBTRACT THE MEAN

Now you have to subtract the mean from the each of the data dimension.This produce the data set whose mean is zero.

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

3.3 CALCULATE THE COVARIANCE MATRIX

For the given 2 dimensional data set we will calculate the covariance matrix of dimension 2X2 (cov(x,x),cov(x,y),cov(y,x),co
The general formula for covariance matrix is given by:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

For the given data 2d dataset we get the get the following covariance matrix:

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

3.4 CALCULATE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX

Since covariance is the square matrix.We need to calculate the eigenvector and eigenvalue for this matrix.

From the defination of eigenvalue and eigenvector:

From the definition of Eigenvalue and Eigenvector:

[*Covariance matrix*].[*Eigenvector*] = [*Eigenvalue*].[*Eigenvector*]

In the previous step for the covariance matrix these are the eigenvalue and eigenvector:

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

It is important to notice that these eigenvectors are both unit eigenvectors ie. their lengths are both 1 and orthogonal to one another.The reason the two Eigenvectors are orthogonal to each other is because the Eigenvectors should be able to span the whole x-y area.

The Eigenvector is the direction of that line, while the eigenvalue is a number that tells us how the data set is spread out on the line which is an Eigenvector.

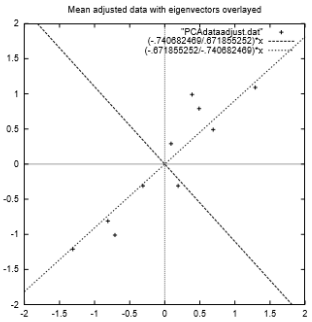

Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

By look at the graph we can deduce the pattern in the data that is how one of the eigenvectors goes through the middle of the points, like drawing a line of best fit. The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

3.5 CHOOSING COMPONENTS AND FORMING A FEATURE VECTOR

once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest.The eigenvector with the highest eigenvalue is the principle component of the data set.This gives you the components in order of significance.Now, if you like, you can decide to ignore the components of lesser significance. You do lose some information, but if the eigenvalues are small, you don't lose much.

Giving the example for the example dataset of 2 dimesion we have 2 eigenvectors,we have two choices either we can have features with both the eigenvectors:

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

Or we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

3.6 CHOOSING COMPONENTS AND FORMING A FEATURE VECTOR

Since the Eigenvectors indicate the direction of the principal components (new axes), we will multiply the original data by the eigenvectors to re-orient our data onto the new axes. This re-oriented data is called a score.

$$FinalData = RowFeatureVector \times RowDataAdjust,$$

where RowFeatureVectoris the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top, and RowDataAdjust is the mean-adjusted data transposed, ie. the data items are in each column, with each row holding a separate dimension.

References

[1] More Math in Latex, Link: https://ctan.math.illinois.edu/info/Math_into_LaTeX-4/Short_Course.pdf

[2] https://medium.com/@dareyadewumi650/understanding-the-role-of-eigenvectors-and-eigenvalues-in-pca-dimensionality-reduction-10186dad0c5c

[3] https://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch18.pdf

[4] https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues