CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS11

Principal Component Analysis (PCA) is often used as a tool in data visualization and reduction of computation load and noise. PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after removing the mean from the data matrix for each feature/column. In this question we will derive PCA. There are four equivalent perspectives to understand PCA. PCA aims to find

1. the Gaussian distribution that best fits with maximum likelihood estimation;
2. the directions of projected maximum variance;
3. the projections of minimum reconstruction error;
4. the best low rank approximation.

In this discussion, we will go through derivations for each of these interpretations, and show how these are all equivalent. First, however, we introduce Rayleigh quotients and present an optimization result that will be extremely useful.

1 Rayleigh Quotients

(a) The Rayleigh quotient is defined as

$$R(M, x) = \frac{x^\top M x}{x^\top x}$$

for a given symmetric matrix $M \in \mathbb{R}^{m \times m}$. What is the interval of possible values of the Rayleigh quotient for a given matrix? Specifically what is

$$\min_x R(M, x) \quad \text{and} \quad \max_x R(M, x)?$$

What values of x attain the bounds?

Solution: For a symmetric matrix, we may consider the spectral decomposition, $M = V\Lambda V^\top$. Assume that

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m.$$

$$\begin{aligned} \frac{x^\top M x}{x^\top x} &= \frac{x^\top V \Lambda V^\top x}{x^\top x} \\ &= \frac{x^\top \sum_{i=1}^m \lambda_i v_i v_i^T x}{x^\top x} \\ &= \frac{x^\top \sum_{i=1}^m \lambda_i v_i v_i^T x}{x^\top x} \end{aligned}$$

Let $y = Vx$, meaning $y_i = v_i^\top x$. Notice that since V is orthogonal,

$$\|y\|_2 = \|Vx\|_2 = \sqrt{x^\top V^\top V x} = \sqrt{x^\top x} = \|x\|_2.$$

We may now apply this substitution in.

$$\begin{aligned}\frac{x^\top M x}{x^\top x} &= \frac{x^\top \sum_{i=1}^m \lambda_i v_i v_i^T x}{x^\top x} \\ &= \frac{\sum_{i=1}^m \lambda_i y_i^2}{\sum_{i=1}^m y_i^2}\end{aligned}$$

Without loss of generality, we can assume that $\sum_{i=1}^m y_i^2 = 1$, or equivalently x and y are unit vectors. We may consider this situation as partitioning 1 into m choices with cost λ_i. The minimum possible cost or value of the Rayleigh quotient is placing all the weight onto the smallest eigenvalue, λ_m. The maximum cost or value of the Rayleigh quotient is placing all the weight onto the largest eigenvalue, λ_1.

From plugging in v_m and v_1, we see that these obtain values λ_m and λ_1. Thus

$$\lambda_m \leq R(M, x) \leq \lambda_1$$

and these values are attained at the corresponding eigenvectors.

(b) How does the Rayleigh quotient relate to the following optimization problems? What does this tell us about the optimum values and the vectors which achieve them? Try to relate these quantities to the singular values and singular vectors of X.

$$\min_{w:\|w\|_2=1} \|Xw\|_2^2 \quad \text{and} \quad \max_{w:\|w\|_2=1} \|Xw\|_2^2.$$

Solution: We may reformulate the minimization problem in the following fashion.

$$\min_{w:\|w\|_2=1} \|Xw\|_2^2 = \min_w \frac{w^\top X^\top X w}{w^\top w}$$

Thus this corresponds to the minimum value of $R(X^\top X, w)$, which is the smallest eigenvalue of $X^\top X$ or the smallest squared singular value of X.

$$\min_{w:\|w\|_2=1} \|Xw\|_2^2 = \min_w R(X^\top X, w) = \sigma_{\min}^2(X).$$

The equivalent holds for the maximum case.

2 The Gaussian MLE Perspective

(a) Assume our data matrix $X \in \mathbb{R}^{n \times d}$ is mean centered. What is the mean and variance of the maximum likelihood estimate for a Gaussian distribution fitting our dataset? What is the covariance matrix when the dataset is centered?

Solution: As we've seen in class, the log likelihood of a Gaussian generating points $X_1, ..., X_n$ is given by

$$\ell(\mu, \Sigma, X_1, ..., X_n) = \ln f(X_1) + \ln f(X_2) + ... + \ln f(X_n) = \frac{nd}{2} \ln \frac{1}{(2\pi)} + \frac{n}{2} \ln |\Sigma^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^\top \Sigma^{-1} (X_i - \mu)$$

Taking the derivative with respect to Σ^{-1}, we get

$$\nabla_\Sigma \ell = \frac{n}{2} \Sigma + \frac{1}{2} (X - \mu)^\top (X - \mu).$$

Setting the derivative equal to 0, we see that the covariance of the MLE for a Gaussian is $\frac{1}{n}(X - \mu)^\top (X - \mu)$.

If the data matrix is mean centered, each feature has mean zero. Thus our Gaussian mean is also the zero vector. Consequently, our covariance matrix is

$$\hat{\Sigma} = \frac{1}{n} X^\top X.$$

(b) Given this Gaussian, how could we construct a k-dimensional basis to project our data, while preserving as much variance as possible?

Solution: The k largest eigenvectors of the covariance matrix $\hat{\Sigma}$ or $X^\top X$, $v_1, \ldots, v_k$ serve as an orthonormal basis of a k dimensional space. These represent the directions with the greatest variance since they correspond to the largest eigenvalues of the covariance matrix. The coordinates in this k dimensional space may be represented as $x^\top v_i$ for $i = 1, \ldots, k$.

3 The Maximum Projected Variance Perspective

(a) We would like to find the vector w such that projecting your data onto w will retain the maximum amount of information (i.e. variance). The projections of our centered data onto w are

$$x_1^\top w, x_2^\top w, \ldots, x_n^\top w,$$

where x_i is the ith row of the matrix X. Compute the mean of and variance of these projections, and show the latter quantity is:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i^\top w\right)^2 = \frac{1}{n}w^\top X^\top X w$$

Solution:

Background: First let us make clear which quantity we are maximizing and what its interpretation is. When we have a set of points $S = \{x_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, what does the term variance even mean? Recall that for random vectors we have the covariance matrix $\Sigma = \mathbb{E}(x - \mathbb{E}(x))(x - \mathbb{E}(x))^\top$. The expectation is taken over the distribution of x. Now there are two questions that arise

- What is the distribution in the case when we have a set of observed samples?
- Given the covariance matrix, what is a scalar variance quantity as a function of that matrix?

With respect to the distribution: On the set of points S we can always define the uniform distribution with $P(x) = \frac{1}{n}$ if $x = x_i$ for some i and zero elsewhere. This is equivalent to the probability of observing x when we draw a random vector from the set S uniformly. This probability mass function corresponds to what we call *the empirical distribution*. This term is especially meaningful when the covariate vectors x are drawn from a true underlying distribution - in which case this empirical distribution is "close" to the underlying one. The covariance matrix of a set of points is taken over this distribution which is thus defined as

$$\Sigma = \mathbb{E}(x - \mathbb{E}(x))(x - \mathbb{E}(x))^\top = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$$

When $\bar{x} = 0$ (i.e. we have subtracted the mean in our samples), we obtain $\Sigma = X^\top X$ (revisit sum of outer product representation of matrix-matrix multiplication if the last step is not clear).

With respect to variance measuring: Given a Gaussian-style probability distribution over x we want to capture the total "amount of randomness" that exists in the system. The quantity $\mathrm{tr}(\Sigma)$ turns out to be a reasonable choice, because of the following fact which has been covered in lecture: If a random vector x has covariance $\Sigma = V\Lambda V^\top$, then $z := V^\top x$ has covariance Λ and all entries of z are independent scalar random variables with z_i having variance λ_i. Since each element of z therefore contributes λ_i noise to the model independently from each other, $\mathrm{tr}\,\Sigma = \sum_{i=1}^n \lambda_i$ represents the total noise introduced. This is the *variance* that we refer to when dealing with sets of points in $d > 1$ dimensions.

Solution: Note that the empirical mean of the projection is

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i^\top w = 0$$

given that our data is centered. Thus, the empirical variance of the projection is

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^\top w-\mu)^2=\frac{1}{n}\sum_{i=1}^{n}(x_i^\top w)^2=\frac{1}{n}w^\top X^\top Xw.$$

(b) We want to find a unit vector w which maximizes this quantity. Formulate this as an optimization problem and find the optimal vector w, along with the corresponding objective value. *Hint: Did we see a similar optimization problem before?*

Solution:

Solution with Rayleigh Quotient: First for either approach, we can ignore the $\frac{1}{n}$ term as it does not affect the maximization. Using the Rayleigh quotient derivation, the maximum value is attained where w is the eigenvector corresponding to the maximum eigenvalue of X^TX or the maximum squared singular value of X.

First-Principles Solution: We start by invoking the spectral decomposition of $X^\top X = V\Lambda V^\top$, which is a symmetric positive semi-definite matrix.

$$\max_{w:\|w\|_2=1} w^\top X^\top Xw = \max_{w:\|w\|_2=1} w^\top V\Lambda V^\top w = \max_{w:\|w\|_2=1} (V^\top w)^\top\Lambda V^\top w \tag{1}$$

Here is an aside: note through this one line proof that left-multiplying a vector by an orthogonal (or rotation) matrix preserves the length of the vector:

$$\|V^\top w\|_2=\sqrt{(V^\top w)^\top(V^\top w)}=\sqrt{w^\top VV^\top w}=\sqrt{w^\top w}=\|w\|_2$$

Define a new variable $z = V^\top w$, and maximize over this variable. Note that because V is invertible, there is a one to one mapping between w and z. Also note that the constraint is the same because the length of the vector w does not change when multiplied by an orthogonal matrix.

$$\max_{z:\|z\|_2=1} z^\top\Lambda z=\max_{z:\|z\|_2=1}\sum_{i=1}^{d}\lambda_i z_i^2$$

From this new formulation, it is obvious to see that we can maximize this by throwing all of our eggs into one basket and setting $z_i^*=1$ if i is the index of the largest eigenvalue, and $z_i^*=0$ otherwise. Thus,

$$z^*=V^\top w^* \implies w^*=Vz^*=v_1$$

where v_1 is the ”principle” eigenvector, and corresponds to λ_1. Plugging this into the objective function, we see that the optimal value is λ_1.

(c) Let us call the solution of the above part w_1. Next, we will use a *greedy procedure* to find the ith component of PCA by doing the following optimization

$$\begin{aligned}\text{maximize}\quad & w_i^\top X^\top Xw_i\\ \text{subject to}\quad & w_i^\top w_i=1\\ & w_i^\top w_j=0\quad\forall j<i,\end{aligned}\tag{2}$$

where $w_j, j < i$ are defined recursively using the same maximization procedure above. Show that the maximizer for this problem is equal to the eigenvector v_i that corresponds to the ith eigenvalue λ_i of matrix $X^\top X$. Also show that optimal value of this problem is equal to λ_i.

Solution: From the previous part, it is obvious to see that we can maximize this by throwing all of our eggs into one basket and setting $z_i^* = 1$ if i is the index of ith largest eigenvalue and others to 0. Plugging this into the objective function, we see that the optimal value is λ_i.

(d) Show that the previous *greedy procedure* finds the global maximum, namely for any $k < d$, $w_1, w_2, \ldots, w_k$ is the solution of the following maximization problem

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^k w_i^\top X^\top X w_i \\ \text{subject to} & w_i^\top w_i = 1 \\ & w_i^\top w_j = 0 \quad \forall i \neq j. \end{array} \tag{3}$$

Solution: It is sufficient to prove that the maximum variance has upper bound $\sum_{i=1}^k \lambda_i$, since this was achieved by the greedy algorithm. For any k orthonormal vectors w_i, variance in this plane is

$$\sum_{i=1}^k w_i^\top X^\top X w_i = \sum_{i=1}^k w_i^\top V^\top \Lambda V w_i = \sum_{j=1}^d \lambda_j \left(\sum_{i=1}^k (V w_i)_j^2 \right)$$

Naturally, we define

$$c_j = \sum_{i=1}^k [V w_i]_j^2 = \sum_{i=1}^k \langle V w_i, e_j \rangle^2,$$

where e_j is the jth standard basis vector corresponding to the jth coordinate. It is sufficient to prove that $c_j \leq 1$ and $\sum_{j=1}^d c_j = k$.

To prove $c_j \leq 1$, note that c_j can equivalently be written as $c_j = \left\|P e_j\right\|_2^2$ where the rows of P consist of $V w_i$. Also note that $V w_1, ..., V w_k$ are orthonormal by the properties of V from the spectral theorem for symmetric matrices, so

$$c_j = \left\|P e_j\right\|_2^2 = e_j^\top P^\top P e_j \leq 1.$$

The second inequality comes from the fact that the maximum eigenvalue of $P^\top P$ is 1.

To prove $\sum_{j=1}^d c_j = k$, we have

$$\sum_{j=1}^d c_j = \sum_{j=1}^d \sum_{i=1}^k \langle V w_i, e_j \rangle^2 = \sum_{i=1}^k \sum_{j=1}^d \langle V w_i, e_j \rangle^2 = \sum_{i=1}^k \|V w_i\|_2^2 = k.$$

Notice that the second to last equality holds by the definition of the squared Euclidean norm — summing up the squares of the coordinates.

We have the constraints that $c_j \leq 1$ and $\sum_{j=1}^d c_j = k$. Thus to maximize $\sum_{j=1}^d \lambda_j c_j$, we must have $c_i = 1$ for $i \in \{1, \ldots, k\}$ and $c_i = 0$ for $i \in \{k+1, \ldots, d\}$.

4 The Minimizing Reconstruction Error Perspective

Our final perspective on PCA is minimizing the perpendicular distance between the principle component subspace and the data points. Let's say we want to find the best 1D space that minimizes the reconstruction error.

(a) Show the (vector) projection of the feature vector x onto the subspace spanned by a unit vector w is

$$P_w(x) = w\left(x^\top w\right). \tag{4}$$

(b) Now, we want to choose w to minimize the reconstruction error. Show that taking w as the minimizer for the corresponding problem below gives us the same result as before.

$$\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i - P_w(x_i)\|_2^2 \tag{5}$$

Solution: We have

$$\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i - P_w(x_i)\|_2^2 \tag{6}$$

$$= \min_{w:|w|=1} \sum_{i=1}^{n} \left(\|x_i\|^2 - 2x_i^\top P_w(x_i) + \|P_w(x_i)\|^2\right) \tag{7}$$

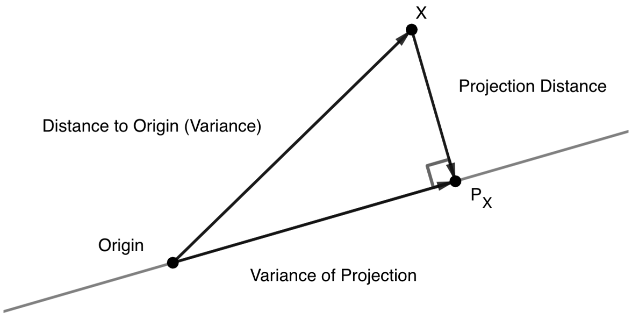
$$= \min_{w:|w|=1} \sum_{i=1}^{n} \left(\|x_i\|^2 - 2(x_i - P_w(x_i))^\top P_w(x_i) - 2P_w(x_i)^\top P_w(x_i) + \|P_w(x_i)\|^2\right) \tag{8}$$

$$= \min_{w:|w|=1} \sum_{i=1}^{n} \left(\|x_i\|^2 - \|P_w(x_i)\|^2\right) \tag{9}$$

$$= \min_{w:|w|=1} \sum_{i=1}^{n} \|x_i\|^2 - \sum_{i=1}^{n} \|(x_i^\top w)w\|^2 \tag{10}$$

$$= \min_{w:|w|=1} \sum_{i=1}^{n} \|x_i\|^2 - \underbrace{\sum_{i=1}^{n} (x_i^\top w)^2}_{\text{Variance Term}}. \tag{11}$$

where the third equality follows from the fact that $P_w(x_i)$ is an orthogonal projection onto the subspace spanned by w (and thus the error is orthogonal to any vector in the subspace including $P_w(x_i)$. Thus we see that minimizing reconstruction error is as same as maximizing variance as what we do in the previous question. Note that this problem can also be shown in alternative ways which you can find in the notes.



The above image serves as a useful visualization. Consider mean centered data. A data point has some fixed distance from the origin. We may consider finding a lower dimensional representation as either maximizing the variance of the projectiong or minimizing the projection distance. The squared quantities must sum to a constant (the distance to the origin or original variance) thus minimizing one is equivalent to maximizing the other.