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?

(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.

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

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?

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

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:

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

(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?*

(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

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

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.

(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

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{k} w_i{}^\top X^\top X w_i \\ \text{subject to} & w_i{}^\top w_i = 1 \\ & w_i{}^\top w_j = 0 \quad \forall i \neq j. \end{array} \tag{2}$$

4 The Minimizing Reconstruction Error Perspective

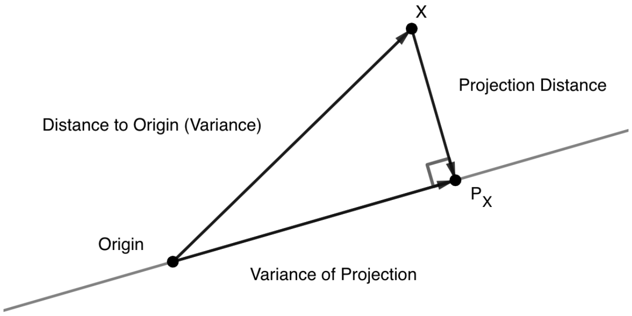
Our final perspective on PCA is minimizing the perpendicular distance between the principle component subspace and the data points. Let's say we want to find the best 1D space that minimizes the reconstruction error.

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

$$P_w(x) = w\left(x^\top w\right). \tag{3}$$

(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.

$$\min_{w:|w|=1} \sum_{i=1}^{n} ||x_i - P_w(x_i)||_2^2 \tag{4}$$



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.