1 PCA Equivalences

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 either 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

given a dataset. In this discussion we will go through derivations for how these are all equivalent.

2 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?

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

The above conclusion tells us range the length input may be modified by the matrix X. This captures some notion of the variance each eigenvector captures. Thus the largest eigenvalue represents the largest amount of variance in one direction. Using the largest eigenvalue direction captures $\frac{\lambda_1}{\sum_{i=1}^m \lambda_i}$ proportion of the total variance. Using the k largest eigenvalue directions captures $\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^m \lambda_i}$ proportion of the total variance.

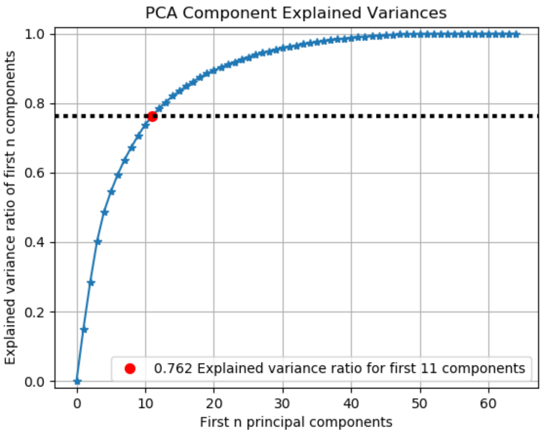


Figure 1: Image from https://scikit-plot.readthedocs.io/en/stable/decomposition.html

The above graph gives an example of a variance plot with respect to the number of principal components used. We can see that using 11 components already captures 76.2% of the variance of the data.

(c) We may consider Rayleigh quotients from an alternate perspective. Consider $R(A,x)$ for an arbitrary x, not necessarily an eigenvector. Show that

$$\arg\min_{\lambda} ||Ax - \lambda x||_2^2 = R(A,x).$$

What happens when x is an eigenvector?

3 Derivation of PCA

(a) Gaussian MLE: 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?

(b) Given this Gaussian, how may we construct a k dimensional basis to project our data?

(c) Maximum Projected Variance: We would like the vector w such that projecting your data onto w will retain the maximum amount of information, i.e., variance. We can formulate the optimization problem as

$$\max_{w:||w||_2=1} \frac{1}{n}\sum_{i=1}^{n}\left(x_i^\top w\right)^2 = \max_{w:||w||_2=1} \frac{1}{n} w^\top X^\top X w \tag{1}$$

where x_i is the feature of ith sample, i.e., the ith row of the matrix X.

Show that the maximizer for this problem is equal to the eigenvector v_1 that corresponds to the largest eigenvalue λ_1 of matrix $X^\top X$. Also show that optimal value of this problem is equal to λ_1/n.

(d) 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

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

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.

(e) 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

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

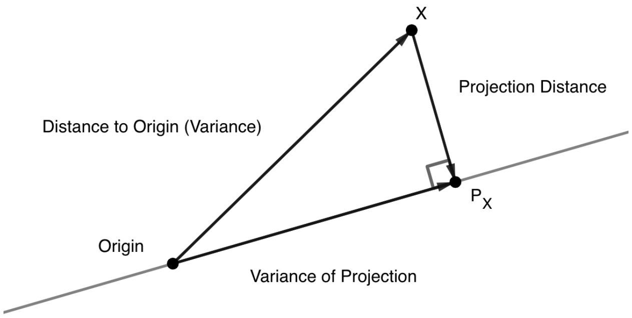
(f) Minimizing Reconstruction Error: Our final perspective on PCA is minimizing the perpendicular distance between the principle component subspace and the data points. Let's say we want to find the best 1D space that minimizes the reconstruction error. The projection of the feature vector x onto the subspace spanned by a unit vector w is

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

Show that the minimizer w for the reconstruction error

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

is as same as the w in Equation (1).



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.

4 Eckart–Young–Mirsky Theorem (Self-Study)

In this problem, we fix an $n \times n$ positive semi-definite matrix A. We will derive, from first principles, the best rank-1 approximation to A. Recall the following metric of approximation: for any integer $1 \leq r \leq n$, we define the best rank-r approximation as any minimizer A_r of

$$\operatorname*{arg\,min}_{M \in R^{n \times n}} \|A - M\|_F : \text{rank}(M) \leq r \tag{6}$$

The notation $\|A\|_F\|$ represents the Frobenius norm. This is equal to the square root of the sum of the squared entries of the matrix, $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$. Note that such a minimizer is not necessarily unique (why?). Since A is positive semi-definite, A_r must be as well: you make take this fact for granted. In this problem, we focus on the special case of $r = 1$. Also assume $A \neq 0$ to avoid any uninteresting, degenerate cases.

(a) Show that $A_1 = mm^T$, where m is any minimizer of the following *unconstrained* optimization problem

$$\min_{m \in R^n} \left\| A - mm^T \right\|_F. \tag{7}$$

(b) Define the function $f : R^n \longrightarrow R$ as $f(m) = \left\| A - mm^T \right\|_F^2$. Compute $\nabla f(m)$.

(c) Argue from part (c) that $A_1 = \lambda_1 v_1 v_1^T$, where λ_1 is the maximum eigenvalue of A and v_1 is a corresponding (unit normalized) eigenvector.

This idea can be generalized to the Eckart–Young–Mirsky Theorem. For a general matrix $A \in \mathbb{R}^{m \times n}$,

$$\operatorname*{arg\,min}_{M:\text{rank}(M) \leq k} \|A - M\|_F = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

where σ_i,u_i,v_i correspond to the singular values, left singular and right singular vectors respectively.

The general SVD formulation for a matrix $X \in \mathbb{R}^{m \times n}$ is $X = U \Sigma V^\top$. In this "full" SVD formulation, we have $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$. Notice that if X is not full rank, then Σ will have zeros along the diagonal, meaning when we expand $U \Sigma V^\top$ some of the columns of U and V are irrelevant.

If $m \geq n$, then this matrix has rank at most n, thus we may consider a formulation where we only consider the first n rows of U and Σ. We may denote these as $U_n \in \mathbb{R}^{m \times n}$, $\Sigma_n \in \mathbb{R}^{n \times n}$. This formulation $U_n \Sigma_n V^\top$ is known as the **thin SVD**.

From the SVD, we can easily recover the best k rank approximation of X, by considering only the first k singular values of Σ, which we may denote as $\Sigma_k \in \mathbb{R}^{m \times k}$. We may also consider the same corresponding r columns of U and V, resulting in $U_k \in \mathbb{R}^{m \times k}$, $V_k \in \mathbb{R}^{n \times k}$. $X_k = U_k \Sigma_k V_k^\top$. This is known as the **truncated SVD**, since we truncate to only include the k dimensions we care about. This is exactly the low rank approximation obtained from Eckart-Young-Mirsky.

If we do not truncate to an arbitrary dimension k and instead reduce to r, the rank of the matrix of X, we do not lose any information. We will still obtain X. $X = U_r \Sigma_r V_r$. This is known as the **compact SVD**.