CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan **DIS12**

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?

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

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

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

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

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

We may now apply this substitution in.

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

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

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

and these values are attained at the corresponding eigenvectors.

(b) How does the Rayleigh quotient relate to the following optimization problems?

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

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

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

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.

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.

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.

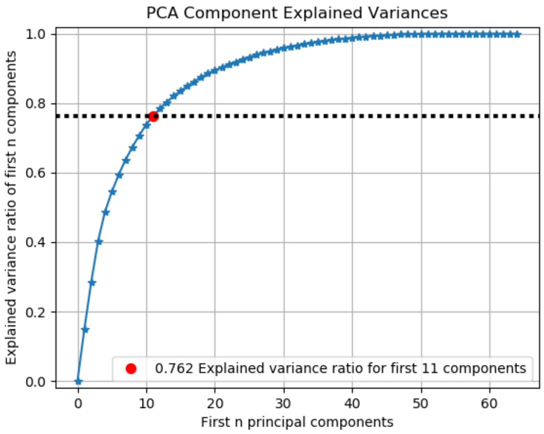


Figure 1: Image from https://scikit-plot.readthedocs.io/en/stable/decomposition.html

(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?

Solution: We may evaluate the minimum by taking the derivative. Let

$$f(\lambda) := \|Ax - \lambda x\|_2^2.$$

$$\begin{aligned} \frac{d}{d\lambda} f(\lambda) &= 0 \\ \frac{d}{d\lambda} \|Ax - \lambda x\|_2^2 &= 0 \\ -2x^\top (Ax - \lambda x) &= 0 \end{aligned}$$

We evaluate the above derivative from the chain rule. From simplifying this, we get

$$x^\top Ax = \lambda x^\top x \implies \lambda = \frac{x^\top Ax}{x^\top x}.$$

When x is an eigenvector, then the optimal value of λ is the corresponding eigenvalue. Thus the rayleigh quotient gives an interpretation as the closest equivalent to an eigenvalue for an arbitrary vector.

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?

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

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

(b) Given this Gaussian, how may we construct a k dimensional basis to project our data? **Solution:** The k largest eigenvectors of the covariance matrix $\hat{\Sigma}$ or $X^\top X$, $v_1, \dots, v_k$ 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 $i = 1, \dots, k$.

(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.

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 rayleight quotient derivation, the maximum value is attained where w is the eigenvector corresponding to the maximum eigenvalue of $X^T X$ or the maximum squared singular value of X.

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

$$\max_{w:||w||_2=1} w^\top X^\top X w = \max_{w:||w||_2=1} w^\top V\Lambda V^\top w = \max_{w:||w||_2=1} (V^\top w)^\top \Lambda V^\top w \tag{2}$$

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:

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

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.

(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

$$\begin{aligned} \text{maximize} \quad & w_i^\top X^\top X w_i \\ \text{subject to} \quad & w_i^\top w_i = 1 \\ & w_i^\top w_j = 0 \quad \forall j < i, \end{aligned} \tag{3}$$

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.

(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

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i=1}^k w_i^\top X^\top X w_i \\ \text{subject to} \quad & w_i^\top w_i = 1 \\ & w_i^\top w_j = 0 \quad \forall i \neq j. \end{aligned} \tag{4}$$

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

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

It is good to notice that the $V w_i$ are themselves orthonormal by the properties of V from the spectral theorem for symmetric matrices. Consequently, it is natural to 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 $c_j \leq 1$, we can think that c_j represents the length of e_j projected to the orthogonal space spanned by $V w_i$, which is always smaller or equal to $||e_j|| = 1$. Let P equal the projection matrix onto the span$\{V w_1, \ldots, V w_k\}$. From properties of projection matrices, we know that P has eigenvalues $\in \{0, 1\}$, and $P^T = P = P^2$. Thus,

$$||P e_j||_2^2 = e_j^\top P^\top P e_j = e_j^\top P e_j \leq \lambda_{\max} P \leq 1.$$

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.

Now we may combine all of the steps together.

$$\begin{aligned}\sum_{i=1}^{k} w_i^\top X^\top X w_i &= \sum_{j=1}^{d} \lambda_j \left(\sum_{i=1}^{k} [Vw_i]_j^2 \right) \\ &= \sum_{j=1}^{d} \lambda_j c_j\end{aligned}$$

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

(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

$$P_w(x) = w\left(x^\top w\right). \tag{5}$$

Show that the minimizer w for the reconstruction error

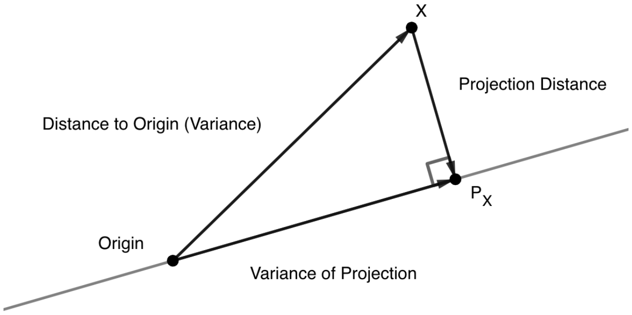
$$\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i - P_w(x_i)\|_2^2 \tag{6}$$

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

Solution: We have

$$\begin{aligned}
&\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i - P_w(x_i)\|_2^2 && (7)\\
=\ &\min_{w:|w|=1} \sum_{i=1}^{n} \left(\|x_i\|^2 - 2x_i^\top P_w(x_i) + \|P_w(x_i)\|^2 \right) && (8)\\
=\ &\min_{w:|w|=1} \sum_{i=1}^{n} \left(\|x_i\|^2 - 2(x_i - P_w(x_i))^\top P_w(x_i) - 2P_w(x_i)^\top P_w(x_i) + \|P_w(x_i)\|^2 \right) && (9)\\
=\ &\min_{w:|w|=1} \sum_{i=1}^{n} \left(\|x_i\|^2 - \|P_w(x_i)\|^2 \right) && (10)\\
=\ &\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i\|^2 - \sum_{i=1}^{n} \|(x_i^\top w) w\|^2 && (11)\\
=\ &\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i\|^2 - \underbrace{\sum_{i=1}^{n} (x_i^\top w)^2}_{\text{Variance Term}}. && (12)
\end{aligned}$$

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 Equation (1). 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.

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

$$\arg\min_{M \in R^{n\times n}} \|A - M\|_F : \text{rank}(M) \leq r \tag{13}$$

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

$$\min_{m \in R^n} \left\|A - mm^T\right\|_F. \tag{14}$$

Solution: Since A is positive semi-definite,

$$\min_{M \in R^{n\times n}} \{\|A - M\|_F : \text{rank}(M) \leq 1\} = \min_{M \in R^{n\times n}} \{\|A - M\|_F : \text{rank}(M) \leq 1, M \geq 0\} \tag{15}$$

$$= \min_{m\in R^n} \left\{\|A - M\|_F : M = mm^T\right\} \tag{16}$$
$$= \min_{m\in R^n} \left\|A - mm^T\right\|_F. \tag{17}$$

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

Solution: For two $n \times n$ matrices A, B, define $\langle A, B\rangle = \mathbf{Tr}(A^T B)$. Now,

$$f(m) = \left\|A - mm^T\right\|_F^2 = \|A\|_F^2 + \left\|mm^T\right\|_F^2 - 2\langle A, mm^T\rangle \tag{18}$$
$$= \|A\|_F^2 + \|m\|_2^4 - 2m^T Am. \tag{19}$$

Recall that $\nabla_m m^T Am = 2Am$. Furthermore, $\nabla_m \|m\|_2^4 = \nabla_m(\|m\|_2^2)^2 = 4mm^T m$. Therefore,

$$\nabla_m f(m) = 4mm^T m - 4Am = 4(mm^T - A)m. \tag{20}$$

(c) Set $\nabla f(m) = 0$ and conclude that all minimizers m of f must satisfy $m = \sqrt{\lambda}v$ where $\lambda \geq 0$ is an eigenvalue of A and v is the corresponding (unit normalized) eigenvector of A.

Solution: Setting $\nabla f(m) = 0$, all minimizers must satisfy the non-linear equation

$$Am = \|m\|_2^2 m. \tag{21}$$

First, we rule out the case that $m = 0$, which satisfies (21). Since $A \neq 0$ and PSD, it must have a non-zero diagonal entry, say A_{kk}. Hence, setting $m = \sqrt{A_{kk}}e_k$, we clearly have $m \neq 0$ and $f(m) < f(0)$.

Therefore, all optimizers of f are non-zero solutions to (21). This means that $m = \alpha v$ for some scalar $\alpha \neq 0$ and eigenvector v with eigenvalue λ of unit norm. Plugging in this parameterization of m into (21) we conclude that $\alpha\lambda v = \alpha^3 v$. Since $\alpha \neq 0$, we can solve for $\alpha = \sqrt{\lambda}$.

(d) 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.

Solution: By part (b), we have at most n candidate solutions for the optimizer, so we simply check them all. Each candidate solution is $M_i = \lambda_i v_i v_i^T$, with $1 \leq i \leq \text{rank}(A)$. However

$$\|A - M_i\|_F^2 = \left\|\sum_{j=1}^r \lambda_j v_j v_j^T - \lambda_i v_i v_i^T\right\|_F^2 = \left\|\sum_{j\neq i}^r \lambda_j v_j v_j^T\right\|_F^2 = \sum_{j\neq i}^r \lambda_j^2 = \|A\|_F^2 - \lambda_i^2. \tag{22}$$

Hence, we have shown that

$$\|A - M_1\|_F^2 \leq \|A - M_i\|_F^2, i = 1, 2, ..., \text{rank}(A). \tag{23}$$

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

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

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**.