mml - ch10

A goal of PCA is to find projections $\tilde{x}_n$ of data points x_n that are as similar to the original data points as possible, with a significantly lower intrinsic dimensionality.

We consider an i.i.d. dataset $\mathcal{X} = \{x_1, \dots x_N\}, x_n \in \mathbb{R}^D$ with mean 0 that posses the data covariance matrix

$$S = \frac{1}{N}\sum_{n=1}^{N} x_n x_n^T.$$ We also assume $\exists$ low-dimensional (compressed rep:

$z_n = B^T x_n \in \mathbb{R}^M$ of x_n, where we define the projection matrix

$B := [b_1, \dots, b_M] \in \mathbb{R}^{D \times M}$. Assume the columns of B are orthonormal.

We want an M-Dimensional subspace $U \subseteq \mathbb{R}^D$, $\dim(U) = M < D$ where we will project the data. projected data is denoted $\tilde{x}_n \in U$, and their coordinates (with respect to the basis vectors $b_1 \dots b_M$ of U) by z_n and the basis vectors $b_1, \dots b_M$.

The projected data should be similar to the original data x_n and minimize the loss due to compression.

Consider $\mathbb{R}^2$, $e_1 = [1,0]$, $e_2 = [0,1]$ Original: $\mathbb{R}^D \ni x \Rightarrow$ Compressed $\mathbb{R}^M \ni z \rightarrow$ Reconstructed $\tilde{x}$

$\begin{bmatrix} 5 \\ 3 \end{bmatrix} = 5e_1 + 3e_2$ now, consider vector of the form $\tilde{x} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \in \mathbb{R}^2$, $z \in \mathbb{R}$. written as: $0e_1 + 2e_2$.

To represent these vectors it is sufficient to remember the coordinate 2 of $\tilde{x}$ with respect to the e_2 vector.

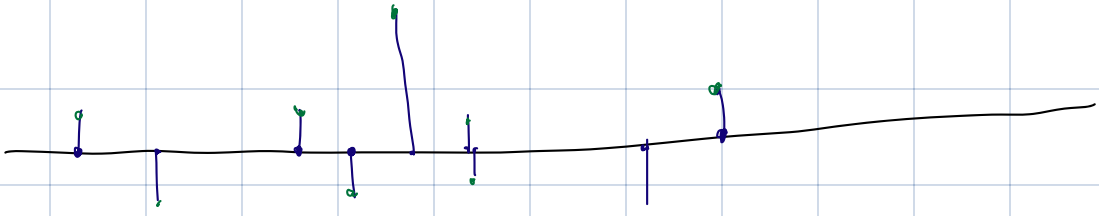
The set of $\tilde{x}$ vectors (+,·) forms a vector subspace U with $\dim(U) = 1 \Rightarrow U = \text{span}[e_2]$

Variance - an indicator of spread in the data PCA - a dimensionality reduction algorithm that maximizes the variance in the low dimensional representation of the data to retain as much information as possible.

Goal: find a matrix B that retains as much information as possible when compressing data by projecting it onto the subspace spanned by columns $b_1 \dots b_M$ of B.

Retaining most information after data compression is equivalent to capturing the largest amount of variance in the low-dimensional code. Code = Coordinates

The variance of low-dimensional code does not depend on the mean of the data. $\Rightarrow$ We assume without loss of generality that the data has mean 0. $\Rightarrow$ the mean of the low-dim code is also 0. $\mathbb{E}_z[z] = \mathbb{E}_x[B^T x] = B^T \mathbb{E}_x[x] = 0$.



PCA finds a lower-dimensional subspace (subtle line) that maintains as much variance (spread of data) as possible when the data (green) is projected onto the subspace (purple)

Direction with Maximal Variance

We start by seeking a single vector $b_1 \in \mathbb{R}^D$ that maximizes the variance of the projected data: aim to maximize the variance of the first coordinate z_1 of $z \in \mathbb{R}^M$: $V_1 := \mathbb{V}[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$ is maximized. The vector b_1 will be the first column of the matrix B and therefore the first of M orthonormal basis vectors that span the lower-dimensional subspace.

The first component of z_n is given by $z_{1n} = b_1^T x_n$, the coordinate of the orthogonal projection of x_n onto the 1D subspace spanned by b_1.

$$V_1 = \frac{1}{N}\sum_{n=1}^{N}(b_1^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1 = b_1^T\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right)b_1 = b_1^T S b_1,$$ where S is the data covariance matrix.

The arbitrarily increasing b_1 increases V_1. A vector b_1 that is twice as long results in a V_1 that is four times as long. $\Rightarrow$ we restrict all solutions to $\|b_1\|^2 = 1$, resulting in a constrained optimization problem in which we seek the direction along which the data varies the most.

b_1 can be found by the constrained optimization problem: $\max_{b_1} b_1^T S b_1$ subject to $\|b_1\|^2 = 1$. We will use the Lagrangian:

$\mathfrak{L}(b_1, \lambda) = b_1^T S b_1 + \lambda_1(1 - b_1^T b_1)$. The partial derivatives of $\mathfrak{L}$ with respect to b_1 and λ_1 are:

$$\frac{\partial \mathfrak{L}}{\partial b_1} = 2b_1^T S - 2\lambda_1 b_1^T, \quad \frac{\partial \mathfrak{L}}{\partial \lambda_1} = 1 - b_1^T b_1. \text{ Setting to 0 gives us relations } \begin{matrix} Sb_1 = \lambda_1 b_1 \\ b_1^T b_1 = 1 \end{matrix}$$

b_1 is an eigenvector of S, and the Lagrange multiplier λ_1 is the eigenvalue: $V_1 = b_1^T S b_1 = \lambda_1 b_1^T b_1 = \lambda_1$,

the variance of the data projected onto a one-dimensional subspace equals the eigenvalue that is associated with the basis vector b_1 that spans this subspace.

The eigenvalue represents the variance of the data when projected onto the one-dimensional subspace.

The quantity $\sqrt{\lambda_1}$ is also called the loading of the unit vector b_1 and represents the standard deviation of the data accounted for by the principal subspace spanned by b_1.

To maximize the variance of the low-dimensional code, we choose the basis vector associated with the largest eigenvalue of S. This eigenvector is called the first principal component.

We determine the effect/contribution of the principal component b_1 in the original data space by mapping z_{1n} back, giving us the projected data point

$\tilde{x}_n = b_1 z_{1n} = b_1 b_1^T x_n \in \mathbb{R}^D$ in the original data space.

Although $\tilde{x}_n$ is a D-dimensional vector, it only requires a single coordinate z_{1n} to represent it with respect to the basis vector $b_1 \in \mathbb{R}^D$.

((0.3.3))

M-dimensional Subspace with Maximal Variance

Assume we have found the first $m-1$ principal components as the $m-1$ eigenvectors of S that were associated with the largest $m-1$ eigenvalues.

S is symmetric, so the spectral theorem states that we can use these eigenvectors to construct an orthonormal eigenbasis of an $(m-1)$-dim subspace of $\mathbb{R}^D$.

We can find the m^{th} principal component can be found by subtracting the effect of the first $m-1$ principal components $b_1, \dots, b_{m-1}$ from the data, trying to find principal components that compress the remaining information.

We have a new data matrix: $\hat{X} := X - \sum_{i=1}^{m-1} b_i b_i^T X = X - B_{m-1}X$, where $X = [x_1, \dots, x_N] \in \mathbb{R}^{D\times N}$ contains the data points as column vectors and $B_{m-1} := \sum_{i=1}^{m-1} b_i b_i^T$ is a projection matrix that projects onto the subspace spanned by $b_1, \dots, b_{m-1}$. → contains the data that has not yet been compressed.

To find the m^{th} principal component, we maximize the variance: $V_m = \mathbb{V}[z_m] = \frac{1}{N}\sum_{n=1}^{N} z_{mn}^2 = \frac{1}{N}\sum_{n=1}^{N}(b_m^T \hat{x}_n)^2 = b_m^T \hat{S} b_m$, subject to constrained optimization $\|b\|^2 = 1$

We find the optimal solution b_m is the eigenvector of $\hat{S}$ that is associated with the largest eigenvalue of $\hat{S}$. b_m is also an eigenvector of $S \Rightarrow$ the set of eigenvectors for both S and $\hat{S}$ are identical.

$\Rightarrow \exists$ D distinct eigenvectors for both S and $\hat{S}$.

Assume we have already found eigenvectors $b_1, \dots, b_{m-1}$ of $\hat{S}$. Consider an eigenvector b_i of S, i.e. $Sb_i = \lambda_i b_i$.

$$\hat{S}b_i = \frac{1}{N}\hat{X}\hat{X}^T b_i = \frac{1}{N}(X - B_{m-1}X)(X - B_{m-1}X)^T b_i = (S - SB_{m-1} - B_{m-1}S + B_{m-1}SB_{m-1})b_i.$$

If $i \geq m$, i.e. b_i is an eigenvector that is not among the first $m-1$ principal components, then b_i is orthogonal to the first $m-1$ principal components and $B_{m-1} b_i = 0$.

If $i < m$, i.e. b_i is among the first $m-1$ principal components, then b_i is a basis vector of the principal subspace onto which B_{m-1} projects.

Since $b_1 \dots b_{m-1}$ are an ONB of this principal subspace, we obtain $B_{m-1} b_i = b_i$.

$B_{m-1} b_i = b_i$ if $i < m$, $B_{m-1} b_i = 0$ if $i \geq m$

Using the case $i \geq m$, we obtain $\hat{S} b_i = (S - B_{m-1} S) b_i = S b_i = \lambda_i b_i$, b_i is also an eigenvector of S with eigenvalue λ_i: $\hat{S} b_m = S b_m = \lambda_m b_m$

b_m is an eigenvector of S and $\hat{S}$ and λ_m is the largest eigenvalue of $\hat{S}$ and the mth largest eigenvalue of S.

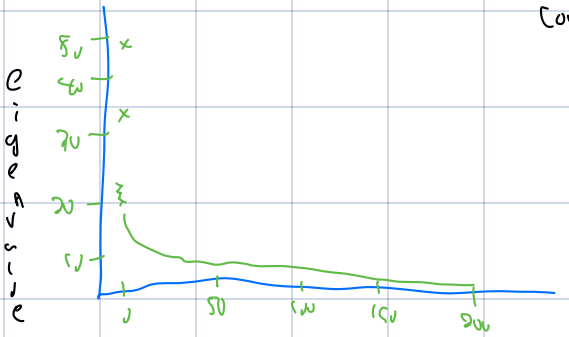
Using the case of $i < m$: $\hat{S} b_i = (S - S B_{m-1} - B_{m-1} S + B_{m-1} S B_{m-1}) b_i = 0 = 0 b_i$

This shows that $b_1 \dots b_{m-1}$ are also eigenvectors of $\hat{S}$, but they are associated with eigenvalue 0 so that $b_1 \dots b_{m-1}$ span the null space of $\hat{S}$.

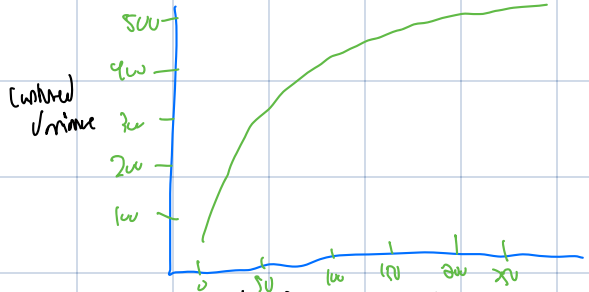
If the eigenvectors of S are part of the $(m-1)$-dimensional principal subspace, then the associated eigenvalue of $\hat{S}$ is 0.

Using $b_m^T b_m = 1$ and $\hat{S} b_m = S b_m = \lambda_m b_m$, the variance of the data projected onto the m^{th} principal components is

$V_m = b_m^T S b_m = \lambda_m b_m^T b_m = \lambda_m$ $\Rightarrow$ when projected onto an M-dimensional subspace, the variance of the data equals the sum of the eigenvalues that are associated with the corresponding eigenvectors of the data covariance matrix.



Eigenvalues of the data covariance matrix of all digits "9" in the MNIST training set.



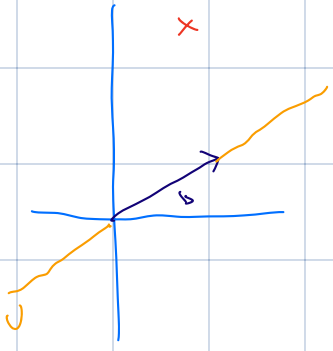
b) Variance captured by the principal components associated with the largest eigenvalues

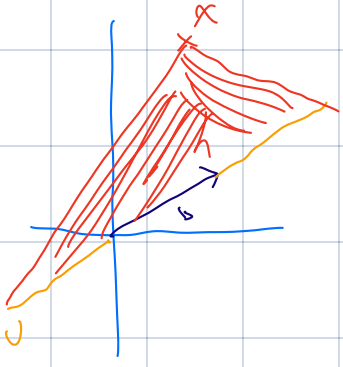
To find an M-dimensional subspace of $\mathbb{R}^D$ that retains as much information as possible, PCA tells us to choose the columns of the matrix B as the M eigenvectors of the data covariance matrix S that are associated with the M largest eigenvalues. The max amount of variance PCA can capture with the first M principal components: $V_M = \sum_{m=1}^{M} \lambda_m$ where the λ_m are the M largest eigenvalues of the data covariance matrix S.

The variance lost by data compression via PCA: $J_M := \sum_{j=M+1}^{D} \lambda_j = V_D - V_M$

the relative variance captured is $\frac{V_M}{V_D}$ and the relative loss is $1 - \frac{V_M}{V_D}$.

(10.3 Projection Perspective)





A projection of a vector x onto a 1-D subspace $U \subseteq \mathbb{R}^2$ spanned by b. Shows difference vectors between x and $\tilde{x}$.

Assume (ONB) $B = (b_1, \dots b_D)$ of $\mathbb{R}^D$. We are interested in finding vectors $\tilde{x} \in \mathbb{R}^D$, which live in lower-dimensional subspace $U \subseteq \mathbb{R}^D$, $\dim(U) = M$, so that

$\tilde{x} = \sum_{m=1}^{M} z_m b_m \in U \subseteq \mathbb{R}^D$ is as similar to x as possible. We want to find the best linear projection onto a lower-dim subspace U of $\mathbb{R}^D$ with $\dim(U) = M$ and orthonormal basis vectors $b_1, \dots b_M$. We call this subspace the principal subspace.

Orthogonal projections are denoted by: $\tilde{x}_n := \sum_{m=1}^{M} z_{mn} b_m = B z_n \in \mathbb{R}^D$, where $z_n := [z_{1n}, \dots, z_{Mn}]^T \in \mathbb{R}^M$ is the coordinate vector of $\tilde{x}_n$ with respect to the basis $(b_1, \dots, b_M)$. We want $\tilde{x}_n$ as similar to x_n as possible.

We will use the Euclidean Norm (squared distance) $\|x - \tilde{x}\|^2$ to measure similarity. Thus, our objective is to minimize the average squared Euclidean distance $J_M := \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2$. To find the optimal linear projection, we need to find the orthonormal basis of the principal subspace and the coordinates $z_n \in \mathbb{R}^M$ of the projections with respect to the basis.

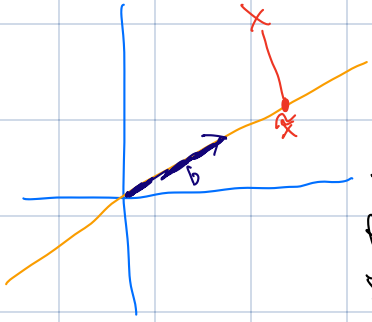
To find the coordinates, there is a two step approach: 1) optimize coordinates z_n for a given ONB $(b_1, \dots b_M)$

2) Find the optimal ONB

Finding the optimal coordinates z corresponds to finding the representation of the linear projection $\tilde{x}$ with respect to b that minimizes the distance between $\tilde{x} - x$.

We assume an ONB $(b_1, \dots, b_M)$ of $U \subseteq \mathbb{R}^D$. We require partial derivatives:

$$\frac{\partial J_M}{\partial z_{in}} = \frac{\partial J_M}{\partial \tilde{x}_n} \frac{\partial \tilde{x}_n}{\partial z_{in}}, \quad \frac{\partial J_M}{\partial \tilde{x}_n} = -\frac{2}{N} (x_n - \tilde{x}_n)^T \in \mathbb{R}^{1 \times D}$$



The vector $\tilde{x}$ that minimizes the distance is it's orthogonal projection onto U. The coordinate of the projection $\tilde{x}$ with respect to the basis vector b that spans U is the factor we need to scale b in order to reach $\tilde{x}$.

$$\frac{\partial \tilde{x}_n}{\partial z_{in}} = \frac{\partial}{\partial z_{in}} \left(\sum_{m=1}^{M} z_{mn} b_m \right) = b_i$$

for $i = 1, \dots M$, we obtain

$$\frac{\partial J_M}{\partial z_{in}} = -\frac{\partial}{N} (x_n - \tilde{x}_n)^T b_i = -\frac{2}{N} \left(x_n - \sum_{m=1}^{M} z_{mn} b_m \right)^T b_i$$

ONB $= -\frac{2}{N} (x_n^T b_i - z_{in} b_i^T b_i) = -\frac{2}{N} (x_n^T b_i - z_{in})$ since $b_i^T b_i = 1$. Setting this partial derivative to zero yields the optimal coordinates:

$z_{in} = x_n^T b_i = b_i^T x_n$. for $i = 1, \dots M$, and $N = 1, \dots N$. $\Rightarrow$ the optimal coordinates of the projection $\tilde{x}_n$ are the coordinates of the orthogonal projection of the original data point x_n onto the one-dimensional subspace spanned by b_i.

- The optimal linear projection $\tilde{x}_n$ of x_n is an orthonormal projection.
- The coordinates of $\tilde{x}_n$ with respect to the basis $(b_1, \dots, b_M)$ are the coordinates of the orthogonal projection of x_n onto the principal subspace.
- An orthogonal projection is the best linear mapping given the objective
- The coordinates z_n of x and z_n of x must be identical for $m = 1, \dots, M$ since $U^\perp = \text{span}[b_{M+1}, \dots b_D]$ is the orthogonal complement of $U = \text{span}[b_1, \dots b_M]$

Finding the Basis of the Principal Subspace

To determine the basis vectors, we rephrase the loss function: $\tilde{x}_n = \sum_{m=1}^{M} z_{mn} b_m = \sum_{m=1}^{M} (x_n^T b_m) b_m$ using symmetry of the dot product,

$$\tilde{x}_n = \left(\sum_{m=1}^{M} b_m b_m^T \right) x_n$$

since we can write x_n as a linear combination of all basis vectors, it holds:

$$x_n = \sum_{d=1}^{D} z_{dn} b_d = \sum_{d=1}^{D} (x_n^T b_d) b_d = \left(\sum_{d=1}^{D} b_d b_d^T \right) x_n = \left(\sum_{m=1}^{M} b_m b_m^T \right) x_n + \left(\sum_{j=M+1}^{D} b_j b_j^T \right) x_n,$$

we split the sum with D terms into a sum over M and a sum over $D-M$ terms. We find that the displacement vector $x_n - \tilde{x}_n$ is:

$$x_n - \tilde{x}_n = \left(\sum_{j=M+1}^{D} b_j b_j^T \right) x_n = \sum_{j=M+1}^{D} (x_n^T b_j) b_j$$

This means that the difference is exactly the projection of the data point onto the orthogonal complement of the principal subspace. The displacement vector $x_n - \tilde{x}_n$ lies in the subspace that is orthogonal to the principal subspace.

PCA finds the best rank-M approximation of the identity matrix

We can now reformulate the loss function: $J_M = \frac{1}{N}\sum_{n=1}^{N}\|x_n - \tilde{x}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N}\left\|\sum_{j=M+1}^{D}(b_j^T x_n)b_j\right\|^2$.

$J_M = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D}(b_j^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} b_j^T x_n b_j^T x_n = \dots = b_j^T x_n x_n^T b_j$

$b_j^T x_n = x_n^T b_j$. We swap sums and obtain: $J_M = \sum_{j=M+1}^{D} b_j^T\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right)b_j = \sum_{j=M+1}^{D} b_j^T S b_j = \sum_{j=M+1}^{D} \mathrm{tr}(b_j^T S b_j) = \sum_{j=M+1}^{D}\mathrm{tr}(S b_j b_j^T) =$

$\mathrm{tr}\left(\left(\sum_{j=M+1}^{D} b_j b_j^T\right)S\right)$

Minimizing the average squared reconstruction error is equivalent to minimizing the projection of the data covariance matrix onto the orthogonal complement of the principal subspace. Minimizing the average squared reconstruction error is equivalent to maximizing the variance of the projected data.

$J_M = \sum_{j=M+1}^{D}\lambda_j$ ← the average squared reconstruction error when projecting onto the M-dimensional principal subspace. where λ_j are the eigenvalues of the data covariance matrix. To minimize the error, we need to select the smallest D−M eigenvalues, implying that their corresponding eigenvectors are the basis of the orthogonal complement of the principal subspace. This means that the basis of the principal subspace comprises the eigenvectors $b_1, \dots, b_M$ that are associated with the largest M eigenvalues of the data covariance matrix.

10.4 Eigenvector Computation

$$S = \frac{1}{N}\sum_{n=1}^{N} x_n x_n^T = \frac{1}{N} X X^T, \quad X = [x_1, \dots, x_N] \in \mathbb{R}^{D\times N}$$

X is a D×N matrix, the transpose of the typical data matrix. To get the eigenvectors/values of S, we can follow two approaches:

- We perform an eigendecomposition and compute the eigenvalues and eigenvectors of S directly.
- We use a singular value decomposition. Since S is symmetric and factorizes into XX^T, the eigenvalues of S are the squared singular values of X.

SVD of X: $\underbrace{X}_{D\times N} = \underbrace{U}_{D\times D}\underbrace{\Sigma}_{D\times N}\underbrace{V^T}_{N\times N}$ where $U \in \mathbb{R}^{D\times D}$ and $V^T \in \mathbb{R}^{N\times N}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{D\times N}$ is a matrix whose only nonzero entries are singular values $\sigma_{ii} \geq 0$. It then follows that

$S = \frac{1}{N} X X^T = \frac{1}{N} U \Sigma \underbrace{V^T V}_{=I_N} \Sigma^T U^T = \frac{1}{N} U \Sigma \Sigma^T U^T$ we get that the columns of U are the eigenvectors of XX^T (and S). The eigenvalues λ_d of S are related to the singular values of X:

$$\lambda_d = \frac{\sigma_d^2}{N}$$

This relationship between the eigenvalues of S and the singular values of X provides the connection between the maximum variance view and the singular value decomposition.

10.6: Key Steps of PCA in Practice

1) Mean Subtraction: We start by centering the data by computing the mean μ of the dataset and subtracting it from every single data point. This ensures that the dataset has mean 0. Mean subtraction is not strictly necessary but reduces the risk of numerical problems.

2) Standardization: Divide the data points by the standard deviation σ_d of the dataset for every dimension $d = 1, \dots, D$.

Now the data is unit-free, and has its variance 1 along each axis.

3) Eigendecomposition of the Covariance Matrix: Compute the data covariance matrix and its eigenvalues and corresponding eigenvectors. Since the covariance matrix is symmetric, the spectral theorem states that we can find an ONB of eigenvectors. The eigenvectors are scaled by the magnitude of the corresponding eigenvalue. The longer vector spans the principal subspace, which we denote by U.

4) Projection: We can project any data point $x_* \in \mathbb{R}^D$ onto the principal subspace: To get this right, we need to standardize x_* using the mean μ_d and standard deviation of the training data in the d^{th} dimension, respectively, so that

$$x_*^{(d)} \leftarrow \frac{x_*^{(d)} - \mu_d}{\sigma_d}, \quad d = 1, \dots, D, \quad \text{where } x_*^{(d)} \text{ is the } d^{th} \text{ component of } x_*.$$

We obtain the projection as $\tilde{x}_* = BB^T x_*$ with coordinates $z_* = B^T x_*$ with respect to the basis of the principal subspace. Here, B is the matrix that contains the eigenvectors that are associated with the largest eigenvalues of the data covariance matrix as columns.

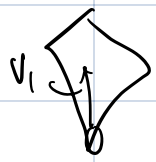
MIT Lin Alg Singular Value Decomposition

$A = U \Sigma V^T$ // Σ diagonal U, V orthogonal

$AV = U\Sigma$ $\quad A = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix}$

Sym pos def

$A = Q \Lambda Q^T$ $\quad \mathbb{R}^n$ row space $\quad \mathbb{R}^m$ column space



$A[v_1 v_2 \dots v_r] = [u_1, u_2, u_r] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \sigma_r & \\ & & & 0 \end{bmatrix}$

row basis vectors

column basis vectors

multiply scalars