Crash Course:
Linear Algebra for Machine Learning

Ben Zhang, Data Science Club

Winter 2019

Purpose

Overview of the important linear algebra concepts required for machine learning.

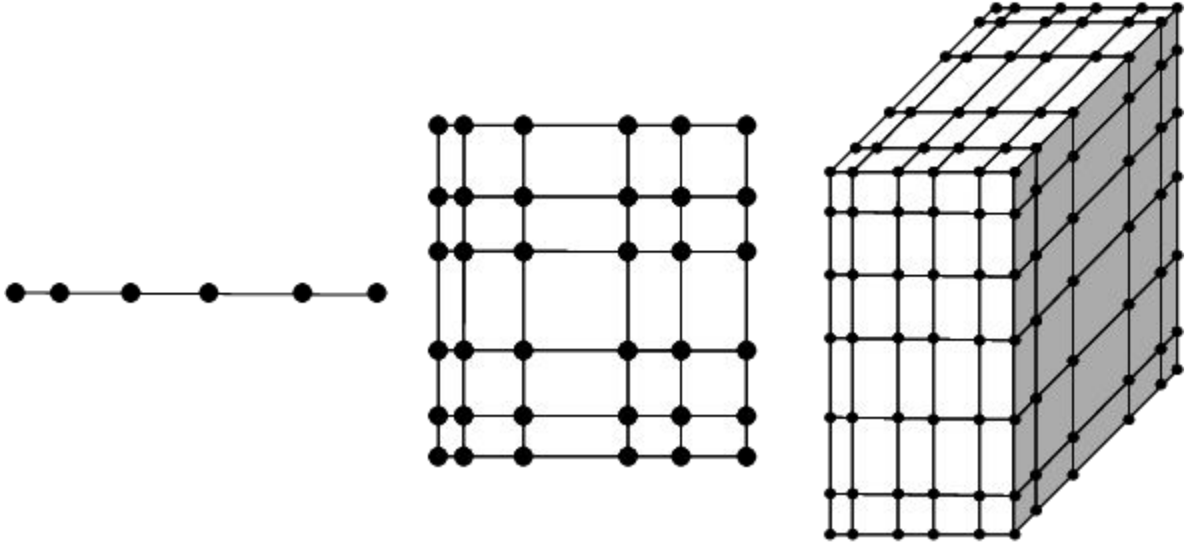
NOT: A proper introduction to Linear Algebra.
See: 3brown1blue, MATH1(3|4)6, MATH2(3|4)5

Part 1: Establishing the basics

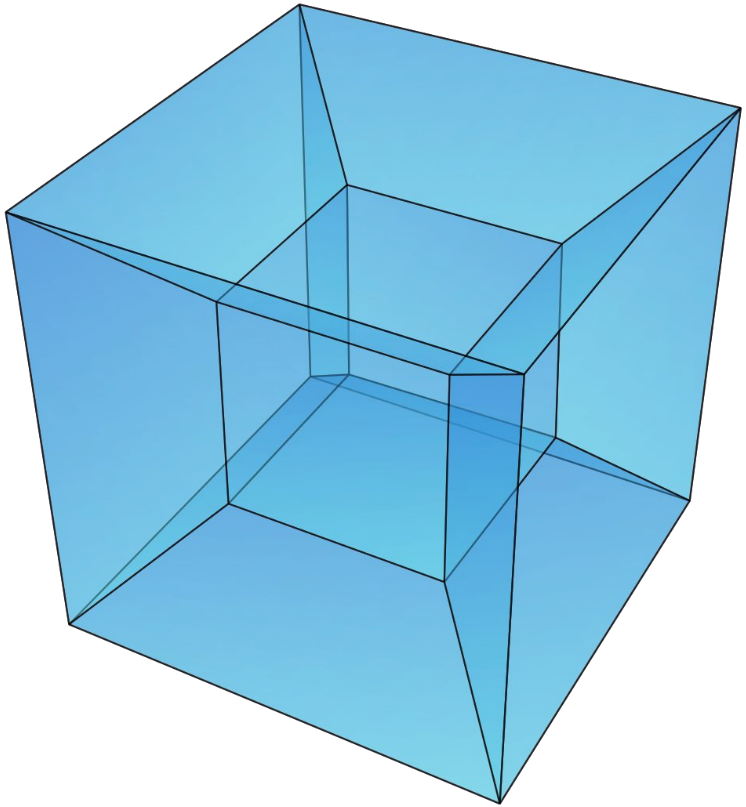
Table of Contents

- Vectors and Matrices
- Norms
- Dot Product
- Matrix Operations
- Transpose and Inverse
- Linear Independence
- Rank
- Determinants
- Eigenvalues

Thinking in more dimensions



The Curse of Dimensionality



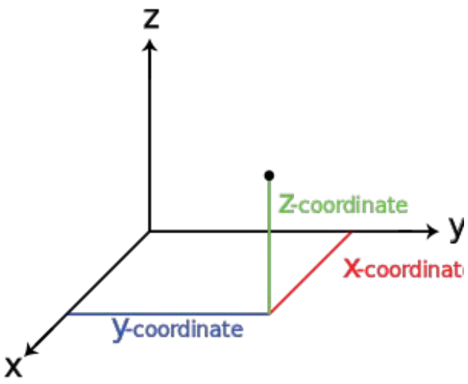
Scalars, Vectors, Matrices

- Scalars are single values. $\mathbb{N}, \mathbb{Q}, \mathbb{R}$
  - Could be from natural numbers, quotients, real numbers.
  - For the most of this lecture, we will use the real numbers.
- Vectors are ordered arrays of values. $\mathbb{R}^n$
  - Indices numbered 1 to n.
  - Column-wise notation, can consider as n by 1 matrix.
- Matrices are 2-D arrays of values.
  - Has height and width - height comes first in notation.
  - Not necessarily square.

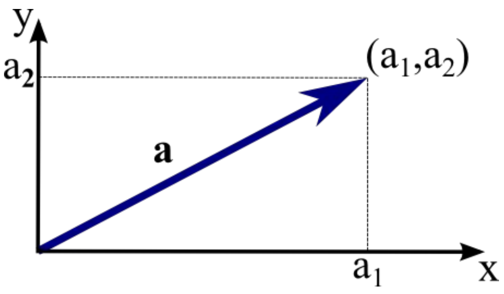
$$\mathbf{x} = \begin{bmatrix} 5 \\ 3 \\ 2 \end{bmatrix} \in \mathbb{R}^3 \qquad x_1 = 5, x_3 = 2$$

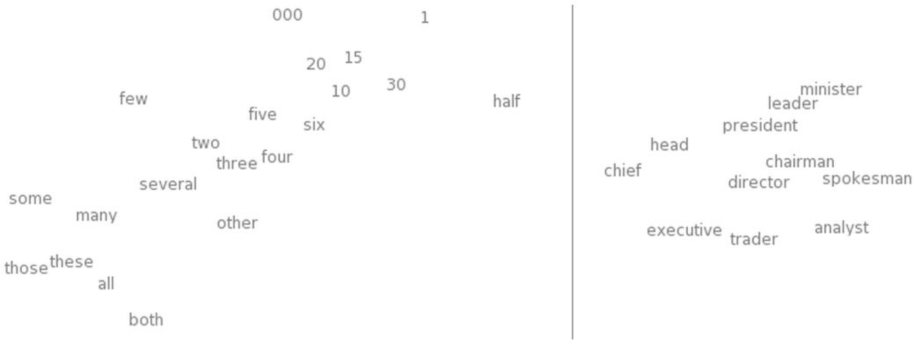
$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ -4 & 1 \\ 8 & -2 \end{bmatrix} \in \mathbb{R}^{3\times 2} \qquad A_{1,2} = 3$$

Intuition on Vectors

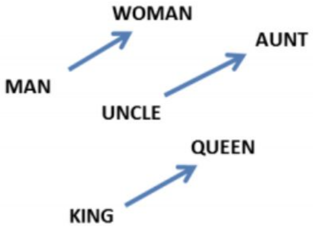


- Interpretation 1: Points in n-dimensional space.
  - Example: word2vec
- Interpretation 2: Linear movement in n-dimensional space
  - Example: word2vec





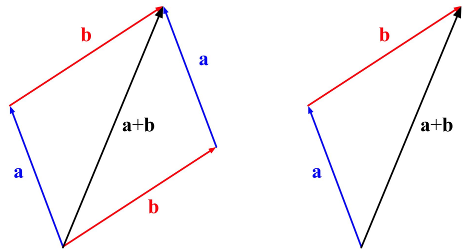
t-SNE visualizations of word embeddings. Left: Number Region; Right: Jobs Region. From Turian *et al.* (2010), see complete image.



From Mikolov *et al.* (2013a)

Addition on Vectors

- Add element-wise.
- Can only add vectors of equal dimensions.
- Associative and commutative.
- Same with matrices.



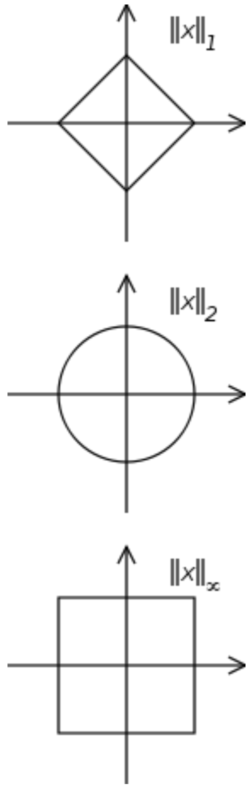
$$\begin{bmatrix} 3 \\ -6 \\ 7 \end{bmatrix} + \begin{bmatrix} -1 \\ 8 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}$$

Norms

- Many different types, serve as a “measure of distance” for vectors.
- Must satisfy the following conditions:

- $f(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0}$
- $f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ (the **triangle inequality**)
- $\forall \alpha \in \mathbb{R}, f(\alpha \boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$

$$||\boldsymbol{x}||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$



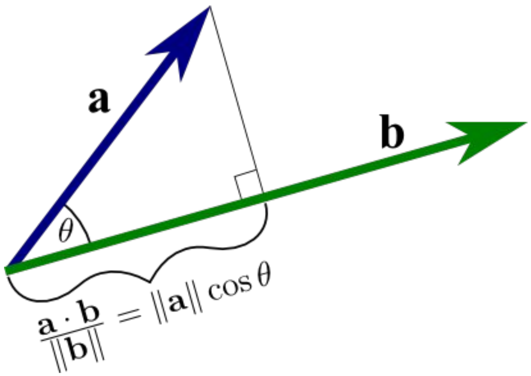
Dot Product

- Takes 2 vectors of the same dimension, returns a scalar.
- A measure of the “alignment” between two vectors, scaled by the lengths.
- Two vectors with dot product zero are **orthogonal** to each other.

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

$$\langle \mathbf{x}, \mathbf{y} \rangle$$

$$\mathbf{x} \cdot \mathbf{y}$$



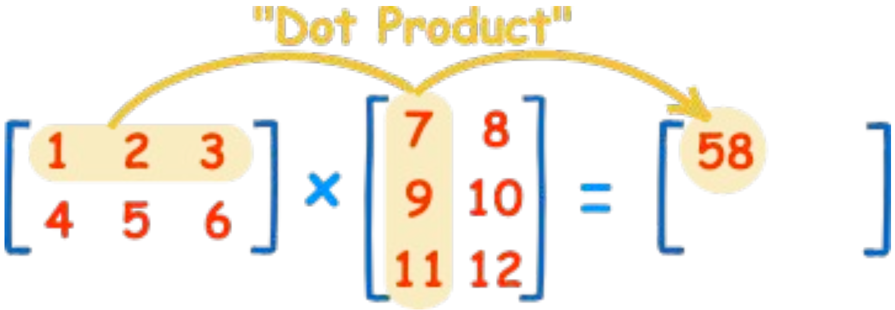
Intuition on Matrices

- Interpretation 1: Ordered collection of vectors (vector of vectors).

- Interpretation 2: Linear transformations on vectors.

Matrix Multiplication

- Multiplying (m, n) matrix with (n, p) matrix yields (m, p) matrix.
- Associative, but not commutative!
- Satisfies the distributive property.
- Identity Matrix, I

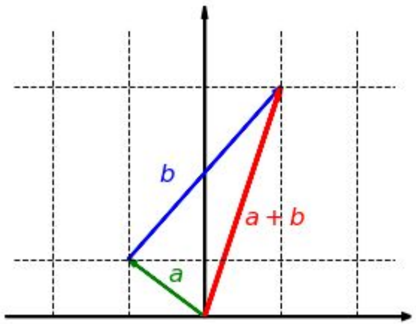


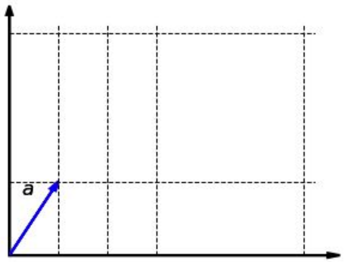
Matrices as linear functions on vectors

- Multiplying a m x n matrix into a n x 1 vector yields a m x 1 vector.
- We can think of this as a linear function from n-dimensional to m-dimensional space.
    - Also need f(0) = 0

$$f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$$

$$f(c\mathbf{u}) = cf(\mathbf{u})$$





$$A \in \mathbb{R}^{m \times n}$$

$$x \in \mathbb{R}^n$$

$$y \in \mathbb{R}^m$$

$$Ax = y$$

$$A \equiv f : \mathbb{R}^n \to \mathbb{R}^m$$

Transpose

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{bmatrix}^{\mathrm{T}}$$

The transpose of a matrix product has a simple form:

$$(\boldsymbol{AB})^{\top} = \boldsymbol{B}^{\top}\boldsymbol{A}^{\top}.$$

Inverse of a Matrix

- Not all matrices are invertible.
  - All invertible matrices are square (dimensional), but not all square matrices are invertible.
  - Square matrices which are not invertible are called **singular**.
  - Singular matrices have determinant 0, which we will not cover.
- Finding inverses is computationally expensive: usually O(n^3)

$$AA^{-1} = A^{-1}A = I$$

Solving systems of linear equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

$$\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

Special Matrices

- Diagonal Matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{bmatrix}$$

- Orthogonal Matrices
  - Orthonormal vectors

$$A^\top A = AA^\top = I.$$

$$A^{-1} = A^\top,$$

- Symmetric Matrices

$$A = A^\top.$$

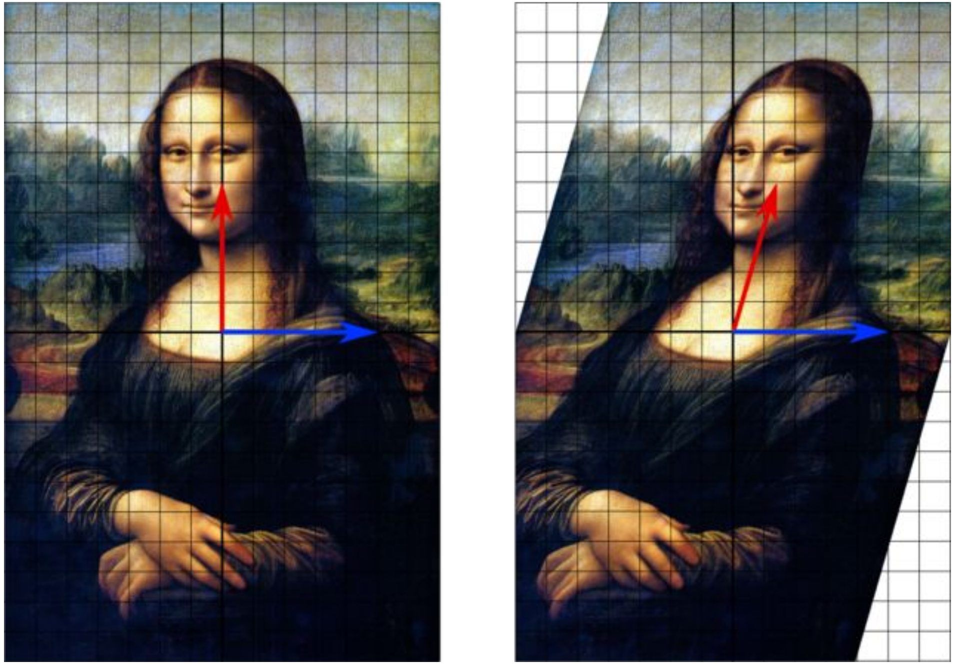
Eigenvectors, Eigenvalues

An **eigenvector** of a square matrix $\boldsymbol{A}$ is a nonzero vector $\boldsymbol{v}$ such that multiplication by $\boldsymbol{A}$ alters only the scale of $\boldsymbol{v}$:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{2.39}$$

The scalar λ is known as the **eigenvalue** corresponding to this eigenvector. (One can also find a **left eigenvector** such that $\boldsymbol{v}^\top \boldsymbol{A} = \lambda \boldsymbol{v}^\top$, but we are usually concerned with right eigenvectors.)

- All scaled eigenvectors are still eigenvectors.
- N by N matrix always has N complex eigenvalues, up to multiplicity
- Symmetric matrices always have N real eigenvalues



Part 2: Applications to ML

Table of Contents

- Eigendecomposition
- Singular Value Decomposition
- Principal Component Analysis

If time permits,

- Linear Regression
- Support Vector Machines

Eigendecomposition

- In the same way that composites can be decomposed into primes, matrices can be decomposed. A must be an n by n matrix.
- Suppose A has n linearly independent eigenvectors, each with an associated eigenvalue.

$$\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$$

$$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$$

$$\boldsymbol{A} = \boldsymbol{V} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{V}^{-1}.$$

Eigendecomposition

- If A is symmetric, there are great properties on the for the eigendecomposition.
- All the eigenvectors are orthonormal, so Q is orthogonal.
- All the eigenvalues are now real.

$$\mathbf{Q} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

$$A = Q\Lambda Q^\top,$$

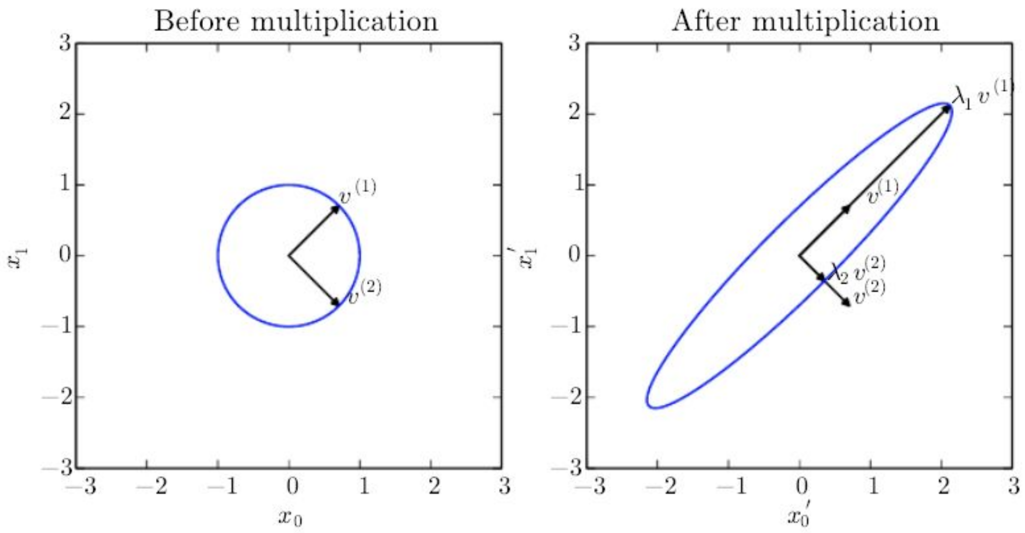


Figure 2.3: An example of the effect of eigenvectors and eigenvalues. Here, we have a matrix $\boldsymbol{A}$ with two orthonormal eigenvectors, $\boldsymbol{v}^{(1)}$ with eigenvalue λ_1 and $\boldsymbol{v}^{(2)}$ with eigenvalue λ_2. *(Left)*We plot the set of all unit vectors $\boldsymbol{u} \in \mathbb{R}^2$ as a unit circle. *(Right)*We plot the set of all points $\boldsymbol{A}\boldsymbol{u}$. By observing the way that $\boldsymbol{A}$ distorts the unit circle, we can see that it scales space in direction $\boldsymbol{v}^{(i)}$ by λ_i.

Useful facts from deriving the Eigenvalues

- A matrix is singular if and only if some eigenvalue is 0
  - The determinant is the product of the eigenvalues
- If any two eigenvectors share the same eigenvalue, then any vector on the span of the eigenvectors is also an eigenvector, with the same eigenvalue.
  - Therefore, even if the eigenvalues are not unique, we can choose a orthogonal set of eigenvectors.
- By convention, we usually sort the eigenvalues from largest to smallest.

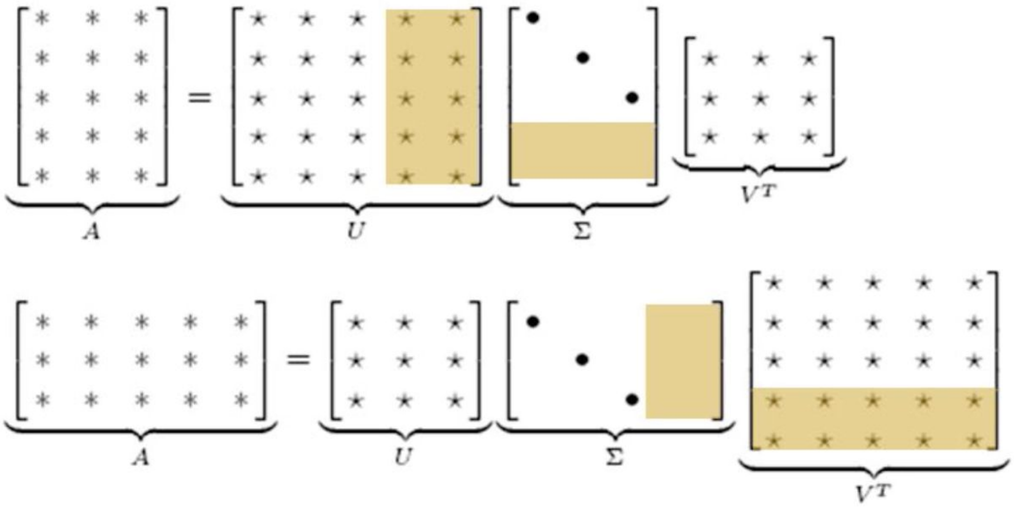
Singular Value Decomposition

- SVD is another way to factorize matrices.
  - Doesn't need the matrix to be a square.
- Every real matrix has a singular value decomposition.

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top. \tag{2.43}$$

Suppose that $\boldsymbol{A}$ is an $m \times n$ matrix. Then $\boldsymbol{U}$ is defined to be an $m \times m$ matrix, $\boldsymbol{D}$ to be an $m \times n$ matrix, and $\boldsymbol{V}$ to be an $n \times n$ matrix.

- Illustration of SVD dimensions and sparseness



Singular Value Decomposition, part 2

$$A = UDV^\top.$$

- U, V are both orthogonal.
- The diagonal values in D are known as the singular values of A.
  - These are the square roots of the eigenvalues of A^T * A.
- Columns of U are the left singular vectors, columns of V are the right singular vectors.

We can actually interpret the singular value decomposition of A in terms of the eigendecomposition of functions of A. The left-singular vectors of A are the eigenvectors of $AA^\top$. The right-singular vectors of A are the eigenvectors of $A^\top A$. The nonzero singular values of A are the square roots of the eigenvalues of $A^\top A$. The same is true for $AA^\top$.

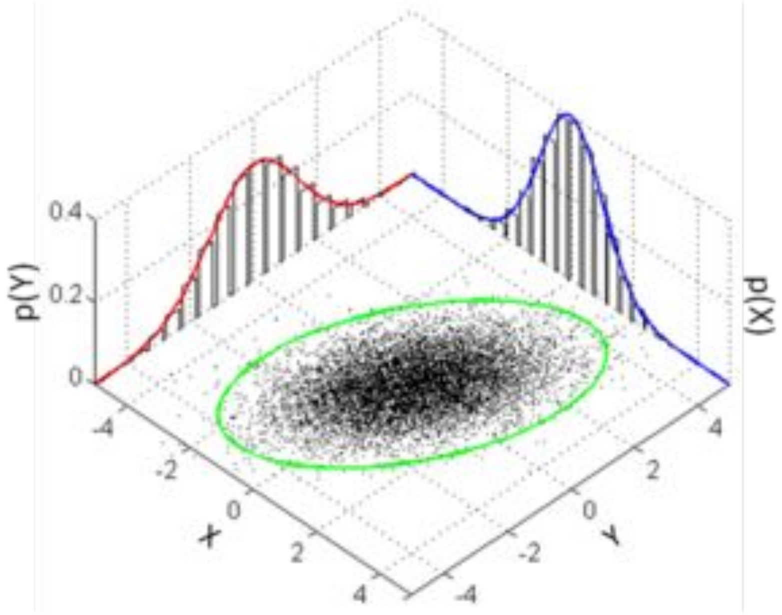
Covariance matrix

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^{\mathrm{T}}$$

$$\mathrm{K}_{X_i X_j} = \mathrm{cov}[X_i, X_j] = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])]$$

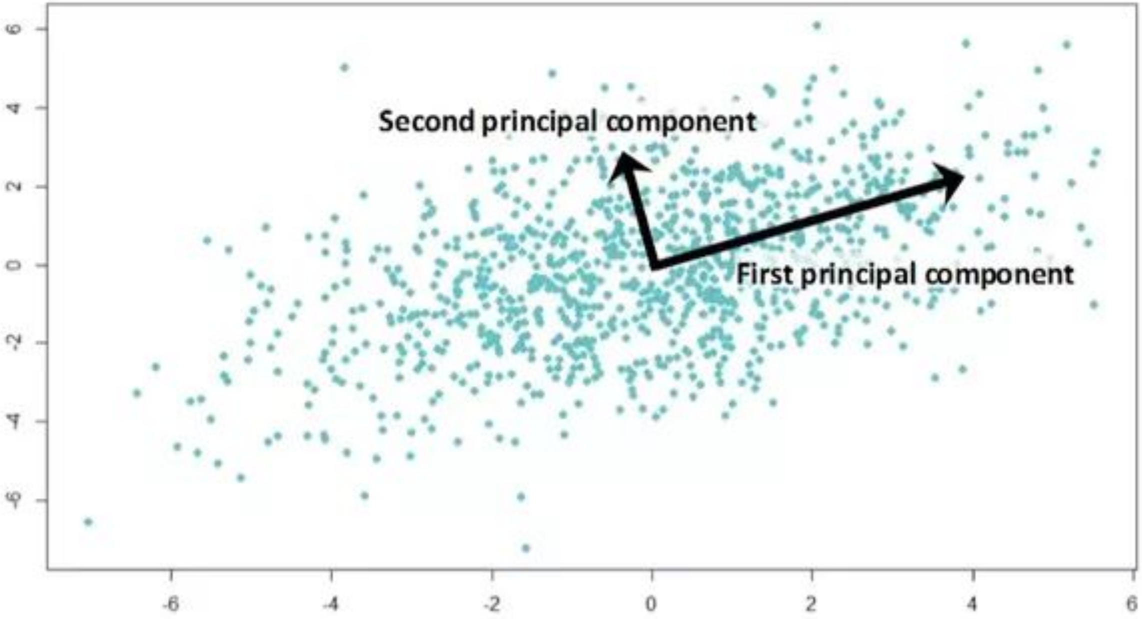
- If each Xi is independent, then the covariance matrix is diagonal.
- Positive Semidefinite: all the eigenvalues are non-negative.
- Covariance matrix written in terms of input data, n by p:

$$\mathbf{C} = \mathbf{X}^\top \mathbf{X}/(n-1)$$

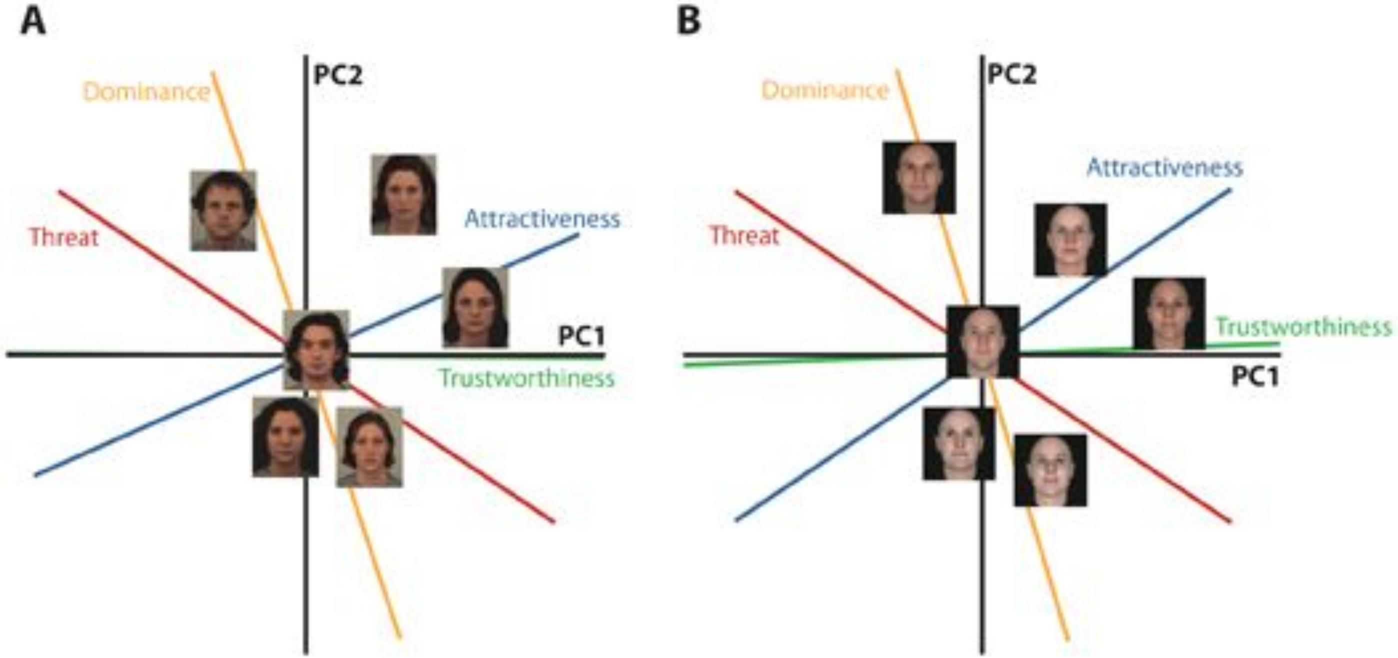


p(Y)
p(X)
0.4
0.2
0
-4
-2
0
2
4
X
Y

Principal Component Analysis



Principal Component Analysis



Principal Component Analysis

- Powerful dimensionality reduction technique.
  - Want to find principal components: low dimensional orthogonal vectors which capture as much variance from the high dimensional data as possible.
  - Want some transform matrix T, which takes high dimensional data and produces low dimensional output.
- Consider input data X, which is n by p matrix.
  - Want eigenvalues and eigenvectors of the **covariance matrix,** ordered by size of eigenvalue.
  - SVD on X: $$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{W}^T$$
- Then,

$$\begin{aligned}\mathbf{T} &= \mathbf{X}\mathbf{W} \\ &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{W}^T\mathbf{W} \\ &= \mathbf{U}\boldsymbol{\Sigma}\end{aligned}$$

Principal Component Analysis (Eigenfaces)

What each principal component looks like

Principal Component Analysis (Eigenfaces pt. 2)

What only the top N principal components looks like

Part 3: Extra Stuff

Measures

- Trace
- Determinant

Other Decompositions

- LU decomposition
- QR decomposition
- Cholesky decomposition

Pseudo-inverse

- Not all matrices have inverses
  - singular matrix
  - non-square matrix
- Moore-Penrose pseudo-inverse is the “closest thing”