Dimensionality Reduction, Eigenvalues and Eigenvectors

Nik Bear Brown

Dimensionality reduction is about converting data of high dimensionality into data of lower dimensionality while keeping most of the information in the data. This allows us to work on larger datasets and identify the data's most relevant features. Anomaly detection (or outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. In this first lesson, we study the theory of dimensionality reduction and introduce essential concepts and jargon, such as eigenvalues, eigenvectors, linear independence, span, vector, scalar, basis of a subspace and linear combination.

Dimensionality Reduction

In machine learning and statistics, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.

Feature selection

Feature selection approaches try to find a subset of the original variables (also called features or attributes). In essence, either using domain knowledge and statisitcal tests to prune away some of the original variables

Feature extraction

Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions.

Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms.

Factor Analysis has was first developed in 1901 by Karl Pearson. Pearson posed a model having one factor that was common across his data:

$$Y_i j = \alpha_i W_1 + \varepsilon_{ij}$$

Or a general multi-factor factor:

$$Y_{ij} = \sum_{k=1}^{K} \alpha_{i,k} W_{j,k} + \varepsilon_{ij}$$

We can use various techniques to estimate $\mathbf{W}_1, \ldots, \mathbf{W}_K$. Choosing k (the number of factors) is a challenge that we'll discuss in further sections.

To illustrate the idea of feature extraction using factors consider the problem of reducing 2D data to 1D.

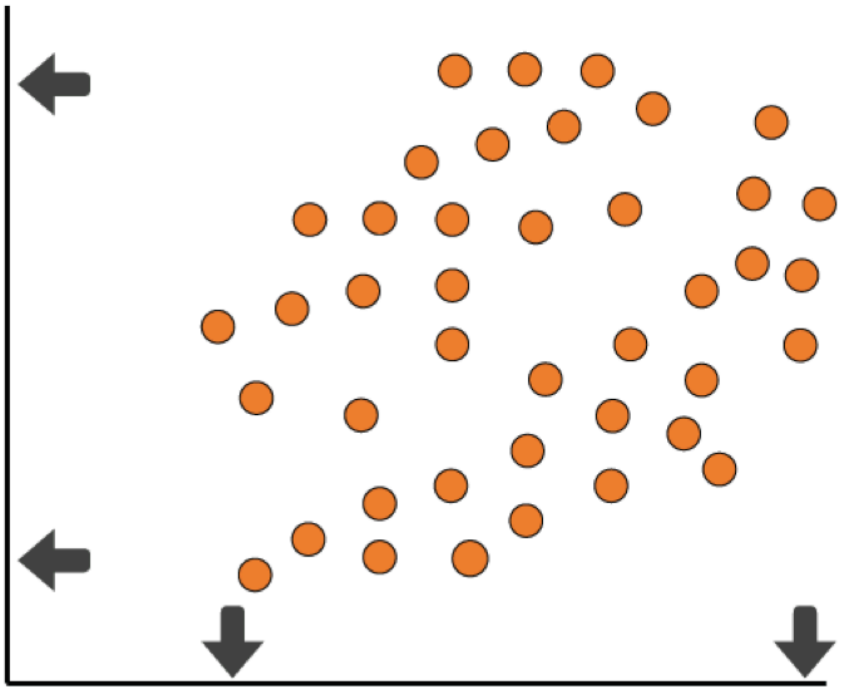

image 2D data to 1D A

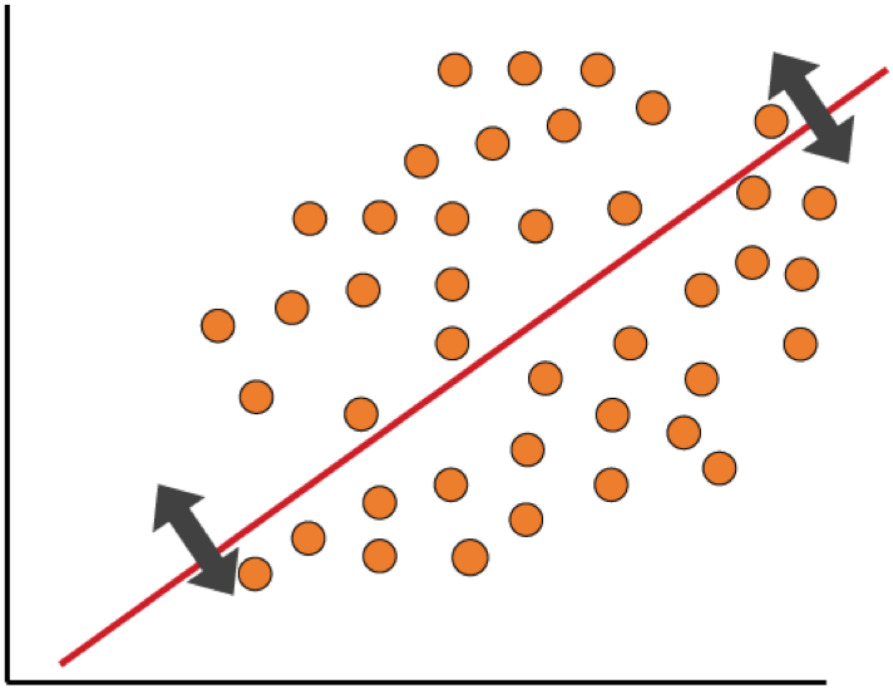

image 2D data to 1D B

Linear Algebra

So how do we find a good basis to project some data?

Some Linear Algebra Jargon

Real coordinate spaces

$$\mathbb{R}^2, \quad \mathbb{R}^3, \ldots, \quad \mathbb{R}^N$$

What is a vector?

- A vector is a quantity having direction as well as magnitude.
- An element of the real coordinate space $\mathbb{R}^N$

Systems of Linear Equations

Linear algebra are used to solve systems of linear equations such as this:

$$\begin{aligned} a + b + c &= 5 \\ 3a - 2b + c &= 3 \\ 2a + b - c &= 1 \end{aligned}$$

We can rewrite and solve this system using matrix algebra notation:

$$\begin{pmatrix} 1 & 1 & 1 \\ 3 & -2 & 1 \\ 2 & 1 & -1 \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 3 & -2 & 1 \\ 2 & 1 & -1 \end{pmatrix}^{-1}\begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix}$$

Multiplying by Vector or Matrix by a Scalar

scalar multiplication of a real Euclidean vector by a positive real number multiplies the magnitude of the vector without changing its direction.

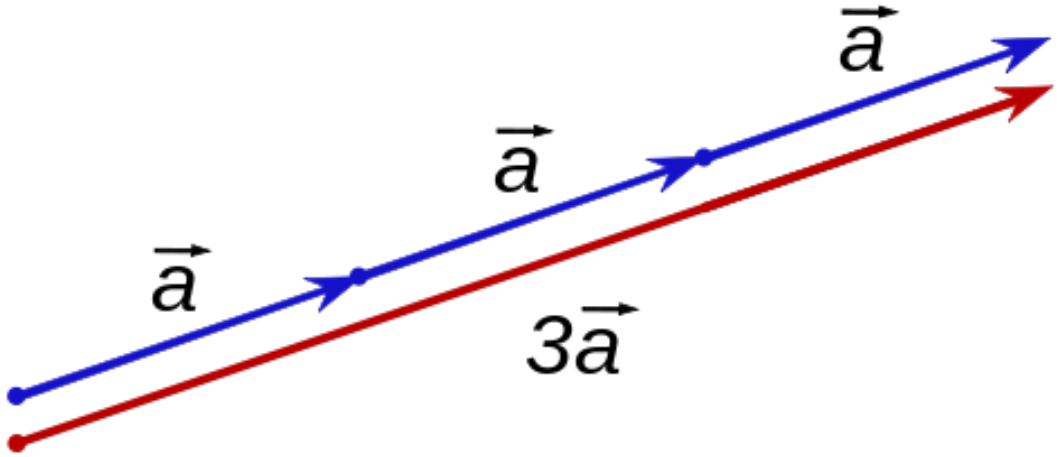

image Scalar multiplication A

Multiplying by a negative value changes its direction.

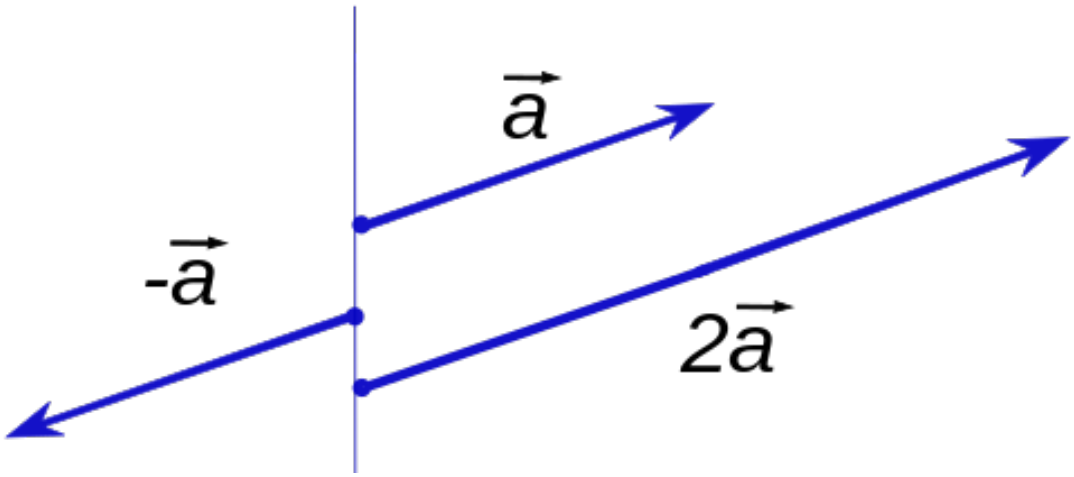

image Scalar multiplication A

If a is scalar and **X** is a matrix then:

$$a\mathbf{X} = \begin{pmatrix} ax_{1,1} & \dots & ax_{1,p} \\ & \vdots & \\ ax_{N,1} & \dots & ax_{N,p} \end{pmatrix}$$

Properties of Scalar Multiplication

Scalar multiplication obeys the following rules:
* Additivity in the scalar: $(c + d)\,\vec{v} = c\,\vec{v} + d\,\vec{v}$;
* Additivity in the vector: $c(\vec{v} + \vec{w}) = c\,\vec{v} + c\,\vec{w}$;
* Compatibility of product of scalars with scalar multiplication: $(cd)\,\vec{v} = c(d\,\vec{v})$;
* Multiplying by 1 does not change a vector: $1\,\vec{v} = \vec{v}$;
* Multiplying by 0 gives the zero vector: $0\,\vec{v} = 0$;
* Multiplying by −1 gives the additive inverse: $(-1)\,\vec{v} = -\vec{v}$.

The Matrix Transpose

The matrix transpose is an operation that changes columns to rows. We use either a $^\top$ to denote transpose.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ x_{2,1} & \dots & x_{2,p} \\ & \vdots & \\ x_{N,1} & \dots & x_{N,p} \end{pmatrix} \Longrightarrow \mathbf{X}^\top = \begin{pmatrix} x_{1,1} & \dots & x_{p,1} \\ x_{1,2} & \dots & x_{p,2} \\ & \vdots & \\ x_{1,N} & \dots & x_{p,N} \end{pmatrix}$$

Matrix multiplication

matrix multiplication is an operation that takes a pair of matrices, and produces another matrix. If A is an n × m matrix and B is an m × p matrix, their matrix product AB is an n × p matrix. The number rows of the first matrix must match the columns of the second.

$$\begin{aligned} a + b + c &= 5 \\ 3a - 2b + c &= 3 \\ 2a + b - c &= 1 \end{aligned}$$

The idea is to multiply the rows of the first matrix by the columns of the second.

$$\begin{pmatrix} 1 & 1 & 1 \\ 3 & -2 & 1 \\ 2 & 1 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} a + b + c \\ 3a - 2b + c \\ 2a + b - c \end{pmatrix}$$

Adding vectors

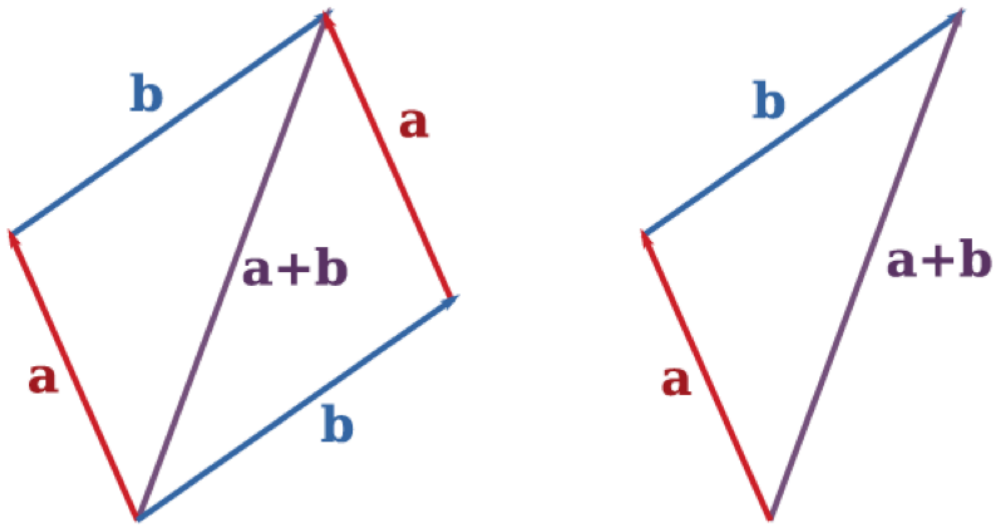

image vector addition A

Unit vector

A unit vector in a normed vector space is a vector of length 1. A unit vector is often denoted by a lowercase letter with a "hat": i-hat (pronounced "i-hat"). The normalized vector or versor û of a non-zero vector u is the unit vector in the direction of u, i.e.,

$$\hat{\mathbf{u}} = \frac{\mathbf{u}}{\| \mathbf{u} \|}$$

u-hat equals the vector u divided by its length where ||u|| is the norm (or length) of u. The term normalized vector is sometimes used as a synonym for unit vector.

from Unit vector - Wikipedia

Linear combinations

We can represent any vector with a linear combination of other vectors. If $\vec{V}$ are vectors $v_1, \ldots, v_n$ and A are scalars $a_1, \ldots, a_n$ then the linear combination of those vectors with those scalars as coefficients is

$$a_1 \vec{v}_1 \quad + \quad a_2 \vec{v}_2 \quad + \quad a_3 \vec{v}_3 \quad + \quad \cdots \quad + \quad a_n \vec{v}_n.$$

For example, consider the vectors $\vec{v}_1 = (1,0,0), \quad \vec{v}_2 = (0,1,0) \quad and \quad \vec{v}_3 = (0,0,1)$. Then any vector in $\mathbb{R}^3$ is a linear combination of $\vec{v}_1$, $\vec{v}_2$ and $\vec{v}_3$. To see that this is so, take an arbitrary vector (a1,a2,a3) in $\mathbb{R}^3$,, and write:

$(a_1, a_2, a_3) = (a_1, 0,0) + (0, a_2, 0) + (0,0, a_3) = a_1(1,0,0) + a_2(0,1,0) + a_3(0,0,1)$
$= a_1 \vec{v}_1 + a_2 \vec{v}_2 + a_3 \vec{v}_3$.

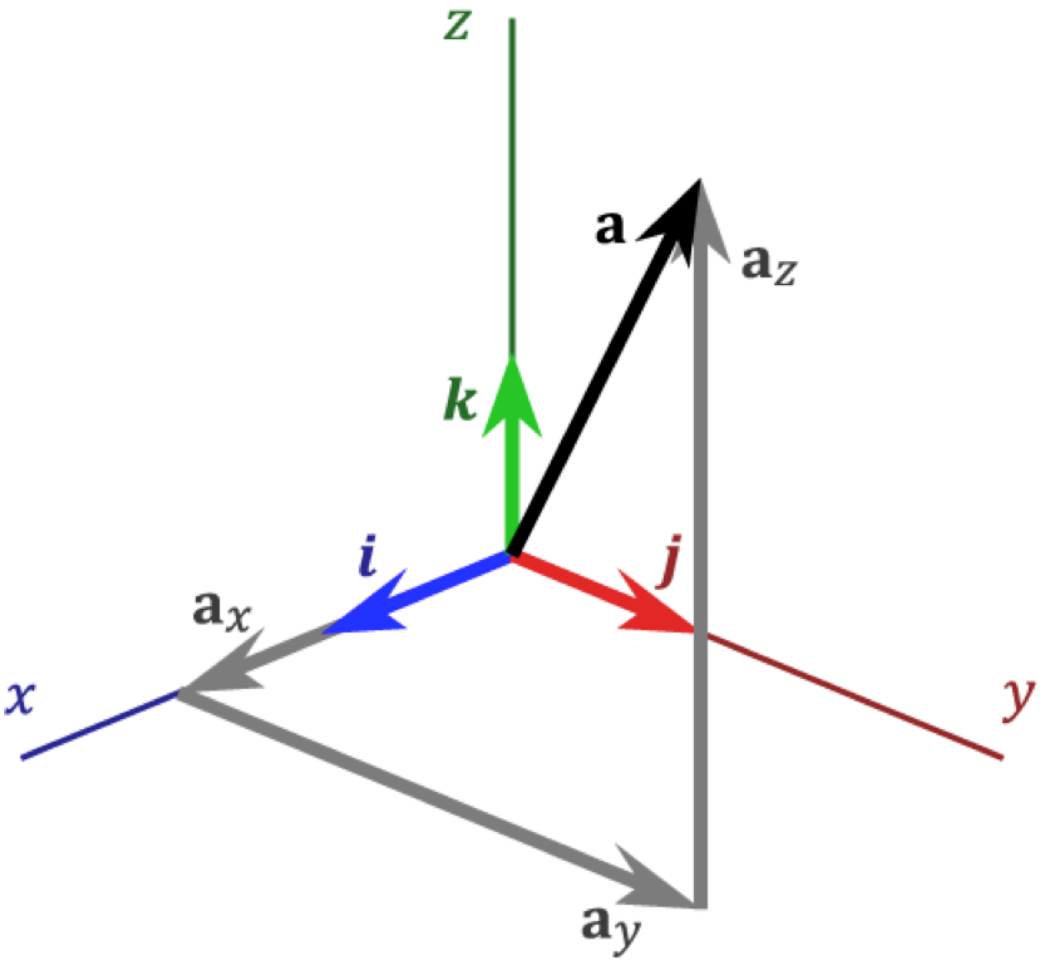

image Linear combination Standard basis A

Linear Spans and Basis

The span of S may be defined as the set of all finite linear combinations of elements of S,

$$\text{span}(S) = \left\{0 + \sum_{i=1}^{k} \lambda_i v_i | k \in \mathbb{N}, v_i \in S, \lambda_i \in \mathbf{K}\right\}$$

A set of vectors in a vector space V is called a basis, or a set of basis vectors, if the vectors are linearly independent and every vector in the vector space is a linear combination of this set. In more general terms, a basis is a linearly independent spanning set.

For example, the real vector space $\mathbb{R}^3$ has {(5,0,0), (0,3,0), (0,0,9)} is a spanning set. This particular spanning set is also a basis.

The set {(1,0,0), (0,1,0), (1,3,0)} is not a spanning set of $\mathbb{R}^3$ as a vector like (1,3,1) cannot be created from this spanning set.

Linear subspaces

A linear subspace (or vector subspace) is a vector space that is a subset of some other (higher-dimension) vector space. For example, $\mathbb{R}^2$ is a subspace of $\mathbb{R}^3$.

Let V be a vector space over the field K, and let W be a subset of V. Then W is a subspace if and only if W satisfies the following three conditions:

- The zero vector, 0, is in W.
- If u and v are elements of W, then the sum u + v is an element of W;
- If u is an element of W and c is a scalar from K, then the product cu is an element of W;

Metric Spaces

A metric space is an ordered pair (M,d) where M is a set and d is a metric on M, i.e., a function * $d: M \times M \to \mathbb{R}$ such that for any * $x, y, z \in M$, the following holds:[1]

- $d(x, y) \geq 0$ (non-negative),
- $d(x, y) = 0 \Leftrightarrow x = y$ (identity of indiscernibles),
- $d(x, y) = d(y, x)$ (symmetry) and
- $d(x,z) d(x,y) + d(y,z) $ (triangle inequality) .

Examples of metric spaces include Euclidean distance, Manhattan distance, Chebyshev distance, the Levenshtein distance and many others.

Vector dot product

Algebraic definition

The dot product of two vectors A = [A1, A2, ..., An] and B = [B1, B2, ..., Bn] is defined as:

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{n} A_i B_i = A_1 B_1 + A_2 B_2 + \cdots + A_n B_n$$

Geometric definition

he magnitude of a vector A is denoted by | |. The dot product of two Euclidean vectors A and B is:

$$\mathbf A \cdot \mathbf B = \left\| \mathbf A \right\| \, \left\| \mathbf B \right\| \cos \theta ,$$

where θ is the angle between A and B.

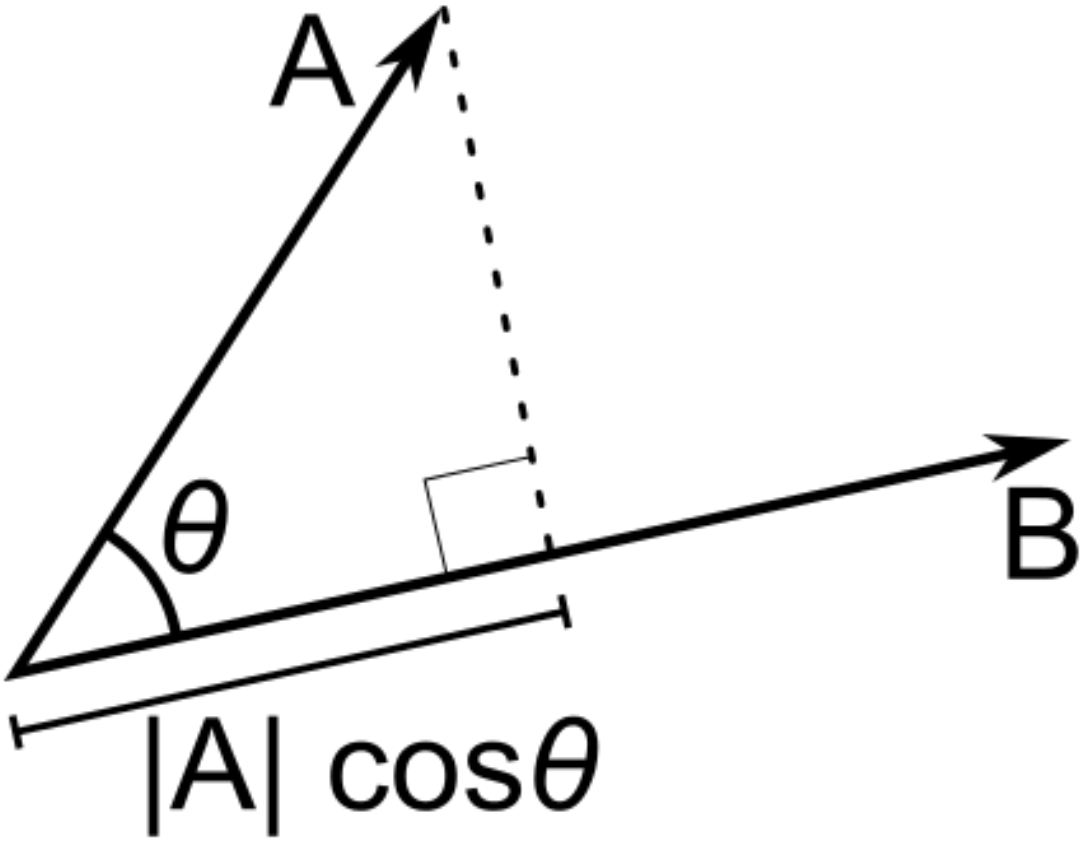

image dot product A

Orthogonal and orthonormal

Two vectors are orthogonal if they are perpendicular, i.e., they form a right angle. Two vectors orthonormal if they are orthogonal and unit vectors. Orthogonal vectors in n-space if their dot product equals zero. That is, if A and B are orthogonal, then the angle between them is 90° and $\mathbf A \cdot \mathbf B = 0 .$

Eigenvalues and eigenvectors

In linear algebra, an eigenvector or characteristic vector of a square matrix is a vector that does not change its direction under the associated linear transformation. In other words—if v is a vector that is not zero, then it is an eigenvector of a square matrix A if Av is a scalar multiple of v. This condition could be written as the equation

$$A\,\vec{v} = \lambda\,\vec{v}$$

where λ is a number (also called a scalar) known as the eigenvalue or characteristic value associated with the eigenvector $\vec{v}$.

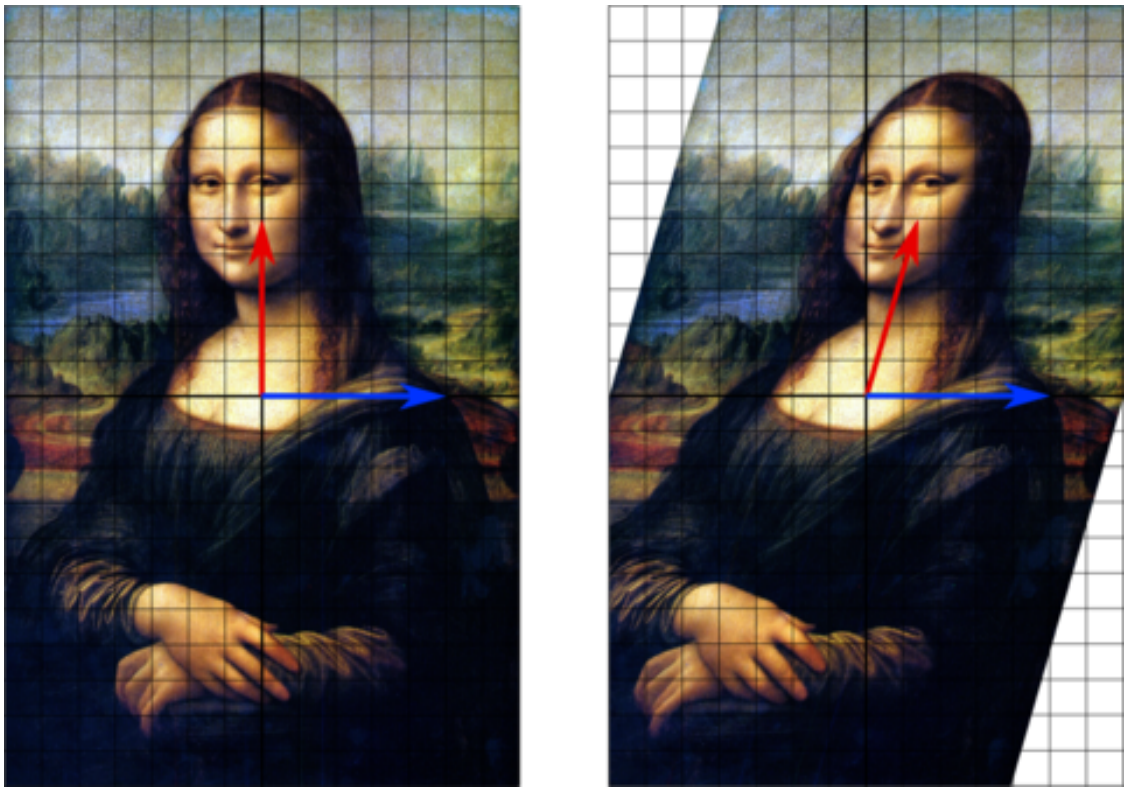

image Mona Lisa eigenvector

In this shear mapping the red arrow changes direction but the blue arrow does not. The blue arrow is an eigenvector of this shear mapping because it doesn't change direction, and since its length is unchanged, its eigenvalue is 1.

For example, Consider n-dimensional vectors that are formed as a list of n real numbers, such as the three dimensional vectors,

$$\mathbf{u} = \begin{Bmatrix} 1 \\ 3 \\ 4 \end{Bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{Bmatrix} -20 \\ -60 \\ -80 \end{Bmatrix}.$$

These vectors are said to be scalar multiples of each other, also parallel or collinear, if there is a scalar λ, such that =.

In this case $$ = −1/20.

Now consider the linear transformation of n-dimensional vectors defined by an n×n matrix A, that is,

$$ A=, or
=

where, for each index i, w_i = A_{i,1} v_1 + A_{i,2} v_2 + + A_{i,n} v_n = *{j = 1}^{n} A*{i,j} v_j.

$$ If it occurs that w and v are scalar multiples, that is if $A\mathbf{v} = \lambda\mathbf{v}$, then v is an eigenvector of the linear transformation A and the scale factor $$ is the eigenvalue corresponding to that eigenvector.

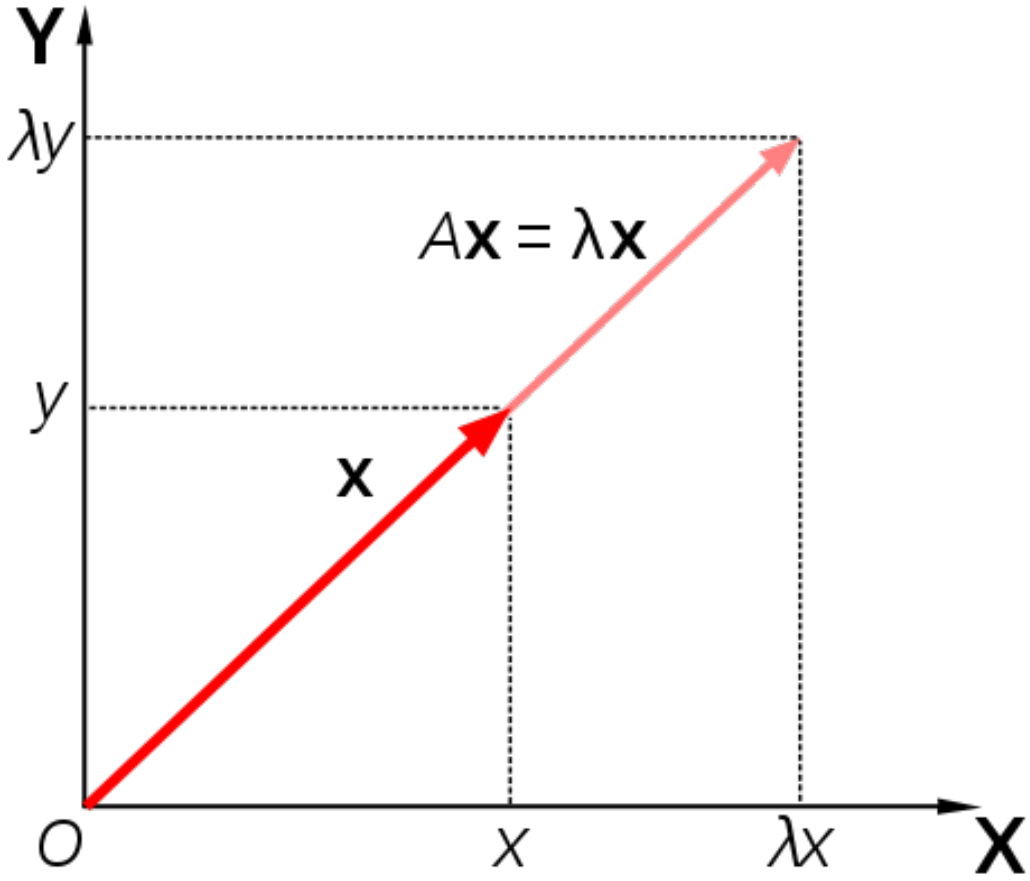

image eigenvector

Matrix A acts by stretching the vector x, not changing its direction, so x is an eigenvector of A.

from Eigenvalues and eigenvectors - Wikipedia

PCA: Principal Component Analyses

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

In Principal component analysis the feature selection is finding a set of linearly uncorrelated vectors (basis selection) of maximal variance. That is, the transformation is defined in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

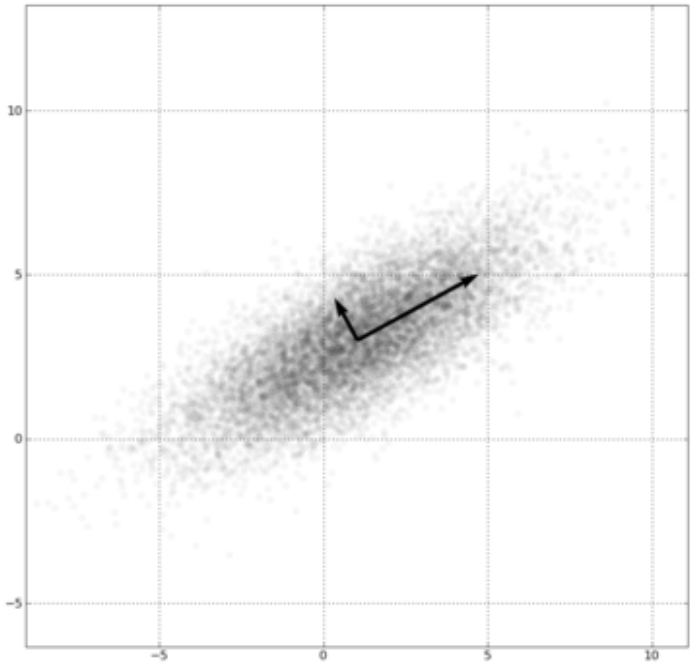

image Principal Component Analyses

First component

The first loading vector $\overrightarrow{w}$ thus has to satisfy

$$ \mathbf{w}_{(1)} = \underset{\Vert \mathbf{w} \Vert = 1}{\operatorname{\arg\,max}}\,\left\{ \sum_i \left(t_1\right)^2_{(i)} \right\} = \underset{\Vert \mathbf{w} \Vert = 1}{\operatorname{\arg\,max}}\,\left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w} \right)^2 \right\} $$

Equivalently, writing this in matrix form gives

$$ \mathbf{w}_{(1)} = \underset{\Vert \mathbf{w} \Vert = 1}{\operatorname{\arg\,max}}\, \{ \Vert \mathbf{Xw} \Vert^2 \} = \underset{\Vert \mathbf{w} \Vert = 1}{\operatorname{\arg\,max}}\, \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X w} \right\} $$

Since $\vec{w}$ has been defined to be a unit vector, it equivalently also satisfies

$$ \mathbf{w}_{(1)} = {\operatorname{\arg\,max}}\, \left\{ \frac{\mathbf{w}^T\mathbf{X}^T \mathbf{X w}}{\mathbf{w}^T \mathbf{w}} \right\} $$

The first component is an eigenvector that maximizes the sum of squares

$$(\mathbf{Y}\mathbf{v}_1)^\top \mathbf{Y}\mathbf{v}_1$$

$\mathbf{v}_1$ is referred to as the *first principal component* (PC). Also referred as *first eigenvector*, $\mathbf{Y}\mathbf{v}_1$ are the projections or coordinates or eigenvalues

Further components

The kth component can be found by subtracting the first k − 1 principal components from X:

$$ \mathbf{\hat{X}}_{k} = \mathbf{X} - \sum_{s = 1}^{k - 1} \mathbf{X} \mathbf{w}_{(s)} \mathbf{w}_{(s)}^{\rm T} $$

and then finding the loading vector which extracts the maximum variance from this new data matrix

$$ \mathbf{w}_{(k)} = \underset{\Vert \mathbf{w} \Vert = 1}{\operatorname{arg\,max}} \left\{ \Vert \mathbf{\hat{X}}_{k} \mathbf{w} \Vert^2 \right\} = {\operatorname{\arg\,max}}\, \left\{ \tfrac{\mathbf{w}^T\mathbf{\hat{X}}_{k}^T \mathbf{\hat{X}}_{k} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\} $$

It turns out that this gives the remaining eigenvectors of $X^T X$, with the maximum values for the quantity in brackets given by their corresponding eigenvalues.

The kth is the vector that

$$\mathbf{v}_k^\top \mathbf{v}_k = 1$$

$$\mathbf{v}_k^\top \mathbf{v}_{k-1} = 0$$

and maximizes

$$(\mathbf{r}\mathbf{v}_k)^\top \mathbf{r}\mathbf{v}_k$$

Properties of Principal Components

- The principal components are eigenvectors
- Maximizes variance in order of the components

- They are orthogonal (and form a basis)
- If use all of the principal components we can get lossless reconstruction
- N diminsions to M diminsions (each diminsion has an eigenvalue the order (highest eigenvalues have highest variance) i.e. can "throw away" the lowest eigenvalues. if eigenvalue has 0 value can throw it away without cost.
- We can readjust to origin 0
- There are very fast algorithms to compute PCA

Nota bene
Note that there may be oroblems of interpretation of the components. Note that a low variance dimension may be important

LDA: Linear Discriminant Analysis

LDA: Linear Discriminant Analysis

An alternative method to calculate eigenvectors from covariance matrix Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.

Singular Value Decomposition

In linear algebra, the singular value decomposition (SVD) is a factorization of a real or complex matrix. A a matrix decomposition or matrix factorization is a factorization of a matrix into a product of matrices.

If $\mathbf{M}$ is a m × n matrix whose entries come from $\mathbb{R}$, then the SVD a factorization, called a singular value decomposition of $\mathbf{M}$, of the form $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}$ gives the factors in columns of $\mathbf{V}$.

Note that,
$\mathbf{D}$ is a m × n diagonal matrix with non-negative real numbers on the diagonal, and $\mathbf{U}$ is an m × m, and $\mathbf{V}$ is an n × n, unitary matrix over $\mathbb{R}$.

The diagonal entries, σ_i, of D are known as the singular values of $\mathbf{M}$. The singular values are usually listed in descending order.

Note that Eigendecomposition is a form of matrix factorization.

$A = VDV^{-1}$, where D is a diagonal matrix formed from the eigenvalues of A, and the columns of V are the corresponding eigenvectors of A.

Differences between Principal Component Analysis and Singular-value decomposition

The singular value decomposition and the eigendecomposition are closely related. PCA viewpoint requires that one compute the eigenvalues and eigenvectors of the

covariance matrix, which is the product $\mathbf{XX^T}$, where $\mathbf{X}$ is the data matrix and SVD constructs the covariance matrix from this decomposition

$$\mathbf{XX^T} = (\mathbf{UDV^T})(\mathbf{UDV^T})^T$$

$$\mathbf{XX^T} = (\mathbf{UDV^T})(\mathbf{VDU^T})$$

and since $\mathbf{V}$ is an orthogonal matrix $\mathbf{V^TV} = \mathbf{I}$),

$$\mathbf{XX^T} = \mathbf{UD^2U^T}$$

That is,
- The left-singular vectors of $\mathbf{M}$ are eigenvectors of $\mathbf{MM}^*$.
- The right-singular vectors of $\mathbf{M}$ are eigenvectors of $\mathbf{M}^*\mathbf{M}$.
- The non-zero singular values of $\mathbf{M}$ (found on the diagonal entries of $\mathbf{D}$) are the square roots of the non-zero eigenvalues of both $\mathbf{M}^*\mathbf{M}$ and $\mathbf{MM}^*$.