CS156 Session 7 - Principal component analysis (PCA)

HCs/LOs	#cs156-unsupervisedlearning
Pre-class	

Readings

(Option 1) Read sections 15.1-15.3 Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge: Cambridge University Press.

> This reading introduces PCA. It explains why we might need to use PCA, and gives the derivation for it. Focus on the high-level understanding, and be able to sketch out the derivation for PCA.

15.1 High-Dimensional Spaces – Low Dimensional Manifolds

- ML problem has high dimensional data: images, bag-of-words, gene expression
- the data does not occupy all spaces in high dimensions

  → linear dimension reductions → **project high dimensional datapoint x onto a lower dimensional vector y**

  $y = Fx + \text{const.}$

  F → dim(y)*dim(x)

15.2 Principal Components Analysis

$$\mathbf{x}^n \approx \mathbf{c} + \sum_{j=1}^{M} y_j^n \mathbf{b}^j \equiv \tilde{\mathbf{x}}^n$$

- bj → the basis vectors of the linear subspace (prinicpal component coefficients or loadings)

15.2.1 Deriving the optimal linear reconstruction

- Find the best basis vector B → minimize the sum of squared difference

$$E(\mathbf{B},\mathbf{Y}) = \sum_{n=1}^{N}\sum_{i=1}^{D}\left[x_i^n - \sum_{j=1}^{M} y_j^n b_i^j\right]^2 = \mathrm{trace}\left((\mathbf{X}-\mathbf{BY})^{\mathsf{T}}(\mathbf{X}-\mathbf{BY})\right)$$

- Optimal solution for B and Y is not unique → need to constrain B to be an orthonormal matrix
- Since we wish to minimise E(B), we therefore **define the basis using the eigenvectors with largest corresponding eigenvalues**

15.2.2 Maximum variance criterion

- To break the invariance of least squares projection with respect to rotations and rescaling, we need an additional criterion
- first searching for the single direction b such that the variance of the data projected onto this direction is maximal amongst all possible such projections

$$y^n = \sum_i b_i x_i^n$$

The projection of a datapoint onto a direction $\mathbf{b}$ is $\mathbf{b}^{\mathsf{T}}\mathbf{x}^n$ for a unit length vector $\mathbf{b}$. Hence the sum of squared projections is

$$\sum_n \left(\mathbf{b}^{\mathsf{T}}\mathbf{x}^n\right)^2 = \mathbf{b}^{\mathsf{T}}\left[\sum_n \mathbf{x}^n(\mathbf{x}^n)^{\mathsf{T}}\right]\mathbf{b} = (N-1)\mathbf{b}^{\mathsf{T}}\mathbf{S}\mathbf{b} \qquad (15.2.15)$$

- the optimal single b which **maximises the projection variance** is given by the eigenvector corresponding to the **largest eigenvalue of S**
- These maximal variance directions found by PCA are called the **principal directions**

15.2.4 PCA and nearest neighbours classification

- For high-dimensional data computing the squared Euclidean distance between vectors can be **expensive**, and also **sensitive to noise**.

  → project the data to a lower dimensional representation first.

$$
\begin{aligned}
(\mathbf{x}^a - \mathbf{x}^b)^\mathsf{T}(\mathbf{x}^a - \mathbf{x}^b) &\approx (\mathbf{E}\mathbf{y}^a + \mathbf{m} - \mathbf{E}\mathbf{y}^b - \mathbf{m})^\mathsf{T}(\mathbf{E}\mathbf{y}^a + \mathbf{m} - \mathbf{E}\mathbf{y}^b - \mathbf{m}) \\
&= (\mathbf{y}^a - \mathbf{y}^b)^\mathsf{T}\mathbf{E}^\mathsf{T}\mathbf{E}(\mathbf{y}^a - \mathbf{y}^b) \\
&= (\mathbf{y}^a - \mathbf{y}^b)^\mathsf{T}(\mathbf{y}^a - \mathbf{y}^b)
\end{aligned}
$$

15.2.5 Comments on PCA

- **The ‘intrinsic’ dimension of data**
  - the reconstruction error is proportional to the sum of the discarded eigenvalues

  → the number of large eigenvalues is the indication of the number of degrees of freedom in the data

  - small eigenvalues are noise
- **Non-linear dimension reduction**

15.3 High Dimensional Data

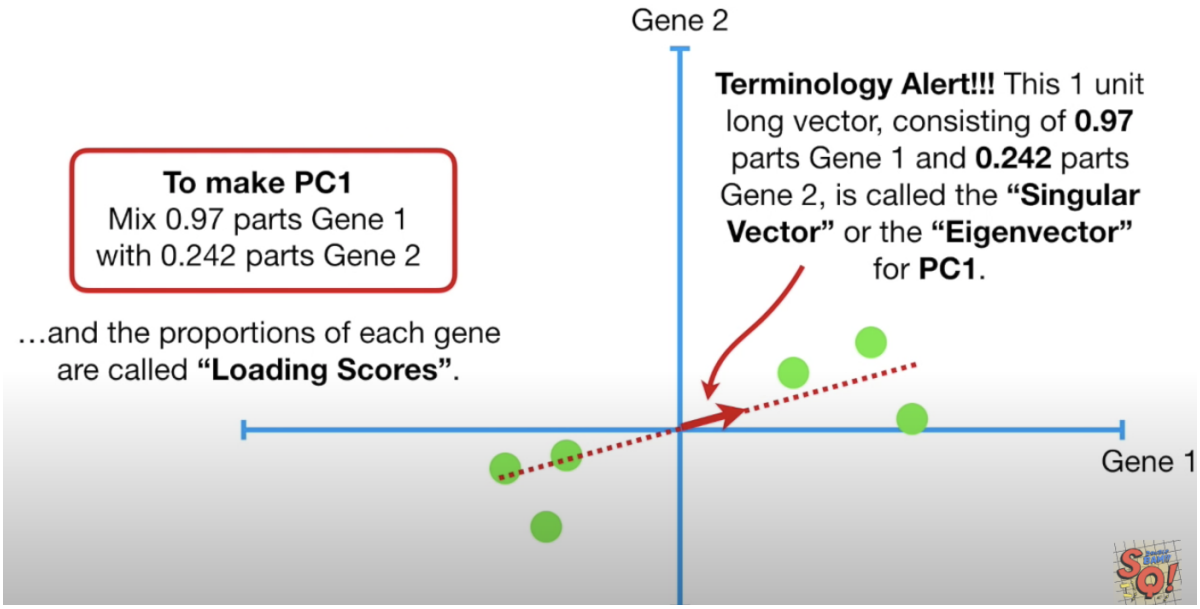
15.3.2 PCA via Singular value decomposition

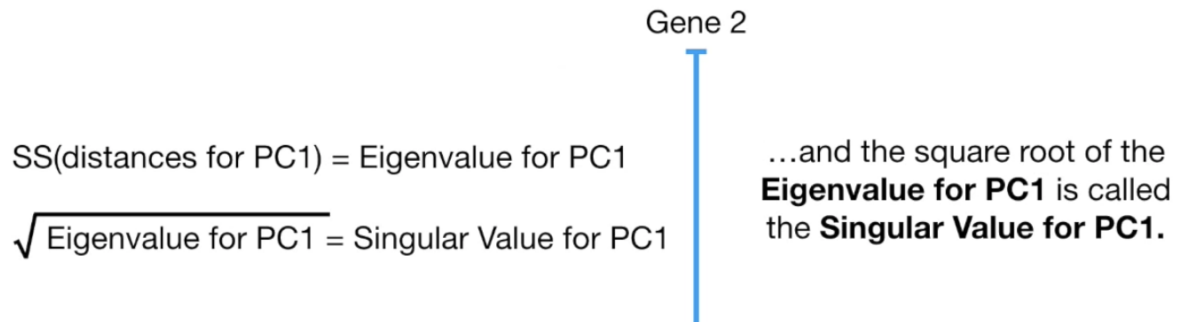
Starmer, J. (2018). StatQuest: Principal Component Analysis (PCA), Step-by-Step Retrieved January 31, 2018, from

https://www.youtube.com/watch?v=FgakZw6K1QQ #### Retrieved from https://www.youtube.com/watch?v=FgakZw6K1QQ

> This video helps to understand PCA visually. Watch this first if you have no prior background.

▼ Notes





$$\frac{SS(\text{distances for PC1})}{n-1} = \text{Variation for PC1}$$

$$\frac{SS(\text{distances for PC2})}{n-1} = \text{Variation for PC2}$$

(Option 2) Read sections 12.2 and 12.3 of Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. Cambridge: The MIT Press.

> Read the following: 1.3.2 Discovering latent factors (p.11) - an introduction to why we need dimensionality reduction 12.2 Principal components analysis (PCA) (p.387) -ignore the first paragraph that references previous material, the sections after 12.2.1 are optional (Proof, Singular value decomposition, Probabilistic PCA, EM algorithm for PCA) 12.2.1 Classical PCA: - statement of the theorem (p.387) 12.3.2 Model selection for PCA (p. 399) - This section introduces metrics to assess how many latent dimensions (components) to choose.

1.3.2 Discovering latent factors

- **latent factors**: the data may appear high dimensional, there may only be a small number of degrees of variability
- **PCA**:
  - an unsupervised version of (multi-output) linear regression
  - observe the high-dimensional response y, but not the low-dimensional "cause" z.

→ Thus the model has the form z → y; we have to "invert the arrow"

→ **infer the latent low-dimensional z from the observed high-dimensional y**

12.2.1 Classical PCA: statement of the theorem

Theorem 12.2.1. *Suppose we want to find an orthogonal set of L linear basis vectors $\mathbf{w}_j \in \mathbb{R}^D$, and the corresponding scores $\mathbf{z}_i \in \mathbb{R}^L$, such that we minimize the average* **reconstruction error**

$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 \tag{12.26}$$

where $\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i$, subject to the constraint that $\mathbf{W}$ is orthonormal. Equivalently, we can write this objective as follows:

$$J(\mathbf{W}, \mathbf{Z}) = ||\mathbf{X} - \mathbf{W}\mathbf{Z}^T||_F^2 \tag{12.27}$$

where $\mathbf{Z}$ *is an* $N \times L$ *matrix with the* $\mathbf{z}_i$ *in its rows, and* $||\mathbf{A}||_F$ *is the* **Frobenius norm** *of matrix* $\mathbf{A}$*, defined by*

$$||\mathbf{A}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\text{tr}(\mathbf{A}^T\mathbf{A})} = ||\mathbf{A}(:)||_2 \tag{12.28}$$

The optimal solution is obtained by setting $\hat{\mathbf{W}} = \mathbf{V}_L$*, where* $\mathbf{V}_L$ *contains the* L *eigenvectors with largest eigenvalues of the empirical covariance matrix,* $\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T$*. (We assume the* $\mathbf{x}_i$ *have zero mean, for notational simplicity.) Furthermore, the optimal low-dimensional encoding of the data is given by* $\hat{\mathbf{z}}_i = \mathbf{W}^T\mathbf{x}_i$*, which is an orthogonal projection of the data onto the column space spanned by the eigenvectors.*

Study Guide

Please consult the study guide given in the course repo:

https://github.com/minerva-schools/cs156

(Don't forget to double check that you are doing the correct session; both the session number and the session title should be the same on Forum as in the repo.)

Pre-class Work

Please consult the pre-class work given in the course repo. Ensure you submit your work to your personal repository in an appropriately-named folder before class starts.

▼ PCA algor concepts

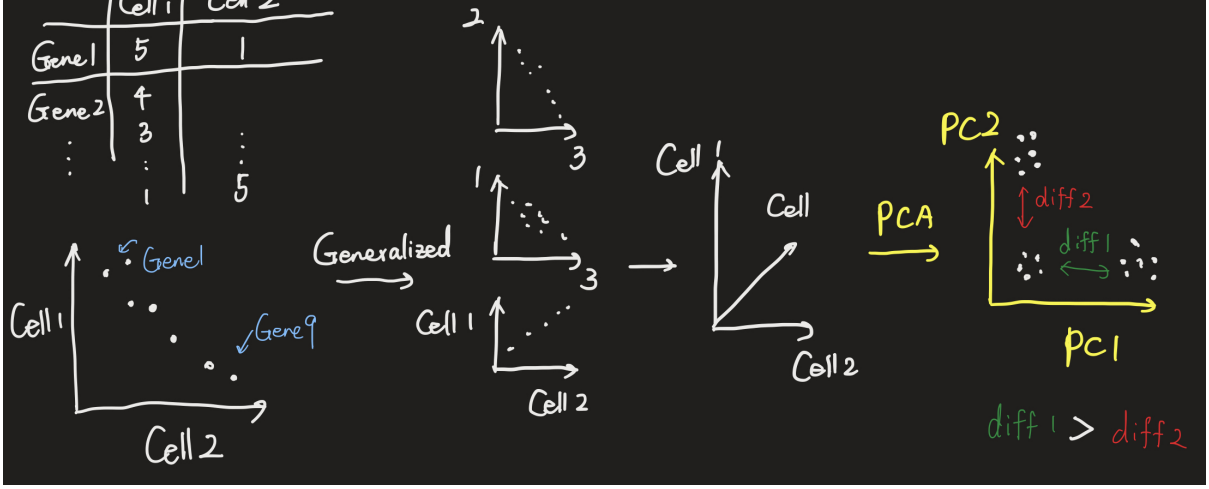
PCA

Saturday, January 29, 2022 2:09 PM

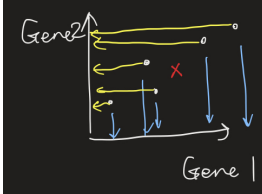
Convert correlations of data into 2D graphs → Dimension reduction

Example:

	Cell 1	Cell 2
Gene 1	5	1
Gene 2	4	
	3	
	⋮	⋮
	1	5



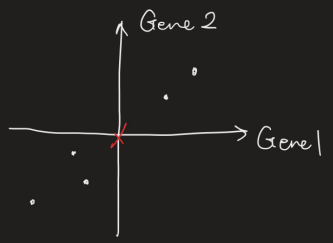
Steps for PCA



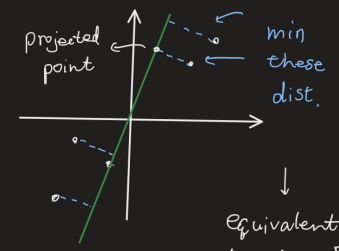
of data distance

① Use the avg value to find the center ✗

② Shift the data → make ✗ locate at (0,0)

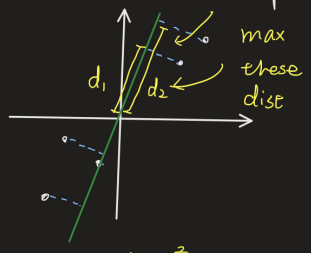


③ Add a line that go through the origin



↓ equivalent based on 畢氏定理

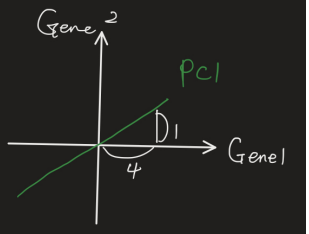
or



$d_1^2 + d_2^2 + d_3^2 + d_4^2$ = SS dist (max this)

④ Find PC1

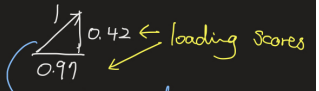
= the line that max SS dist.



PC1 = 4 Gene1 + 1 Gene2

linear combination

↓ Rescale

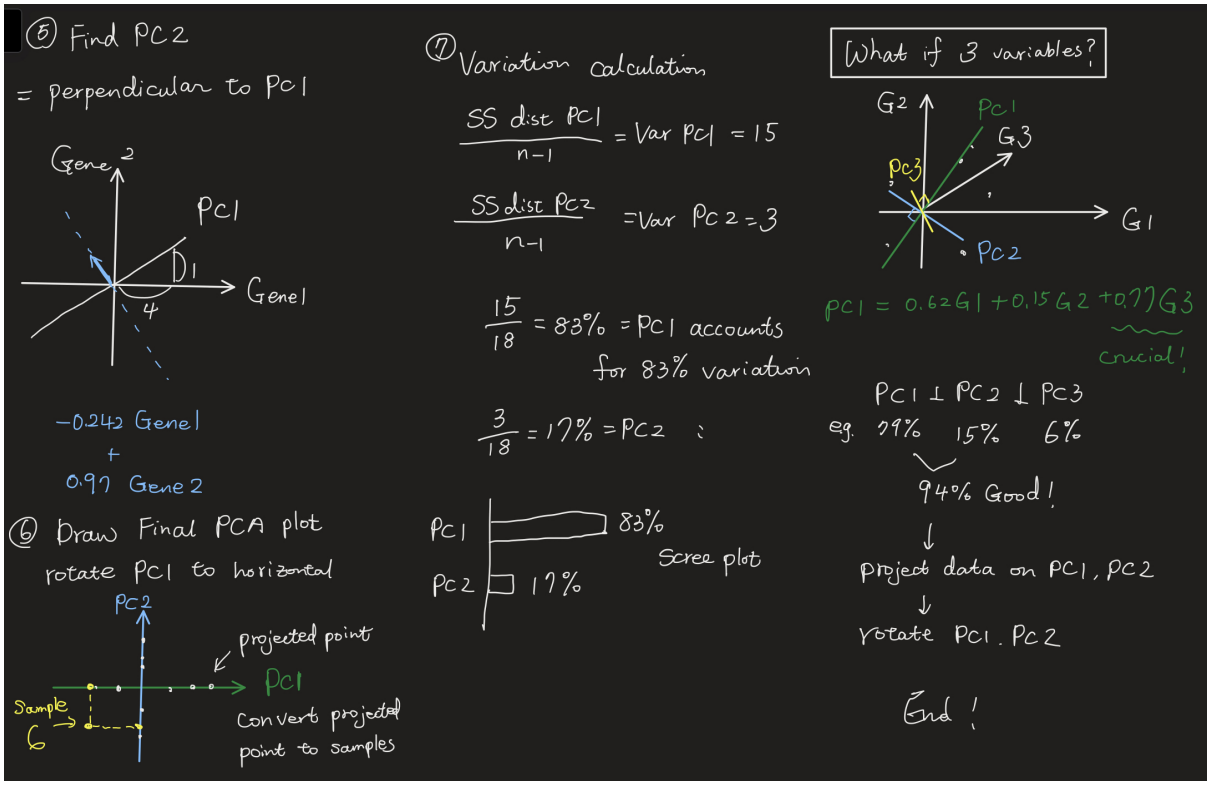


← loading scores

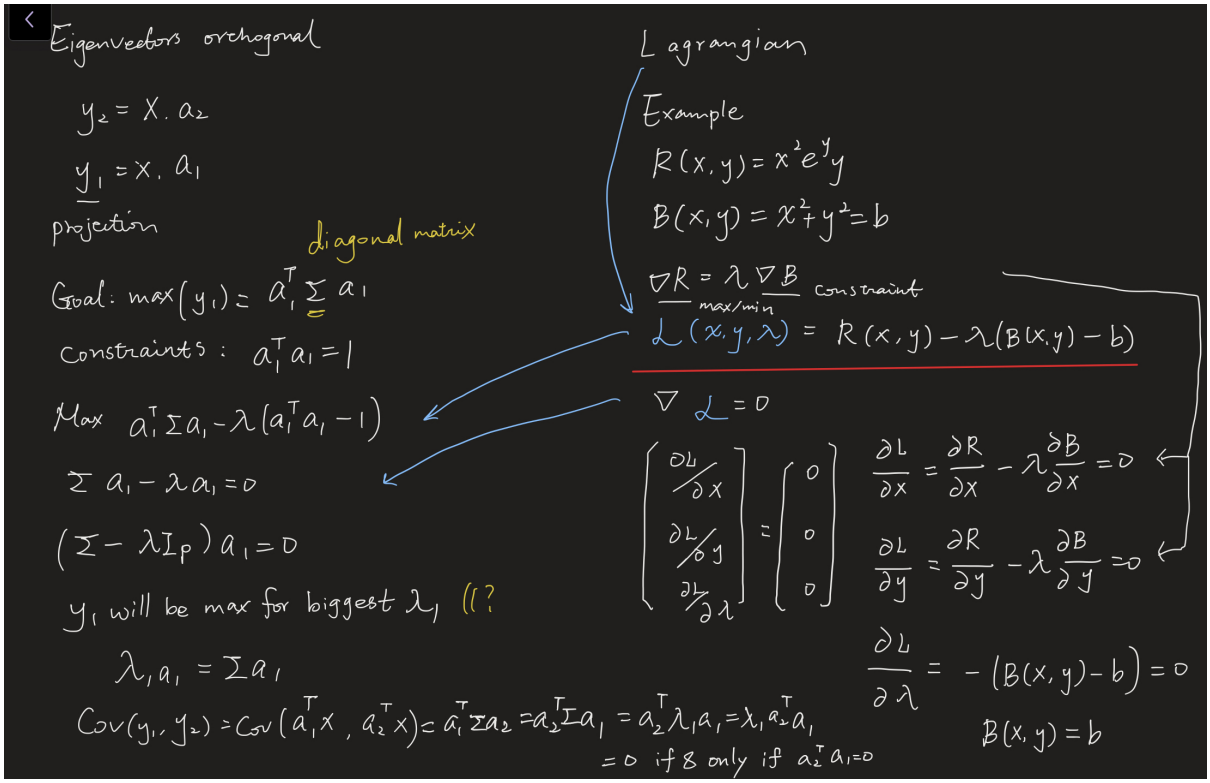
eigenvector

SS dist = eigenvalue

$\sqrt{\text{eigenvalue}}$ = singular value



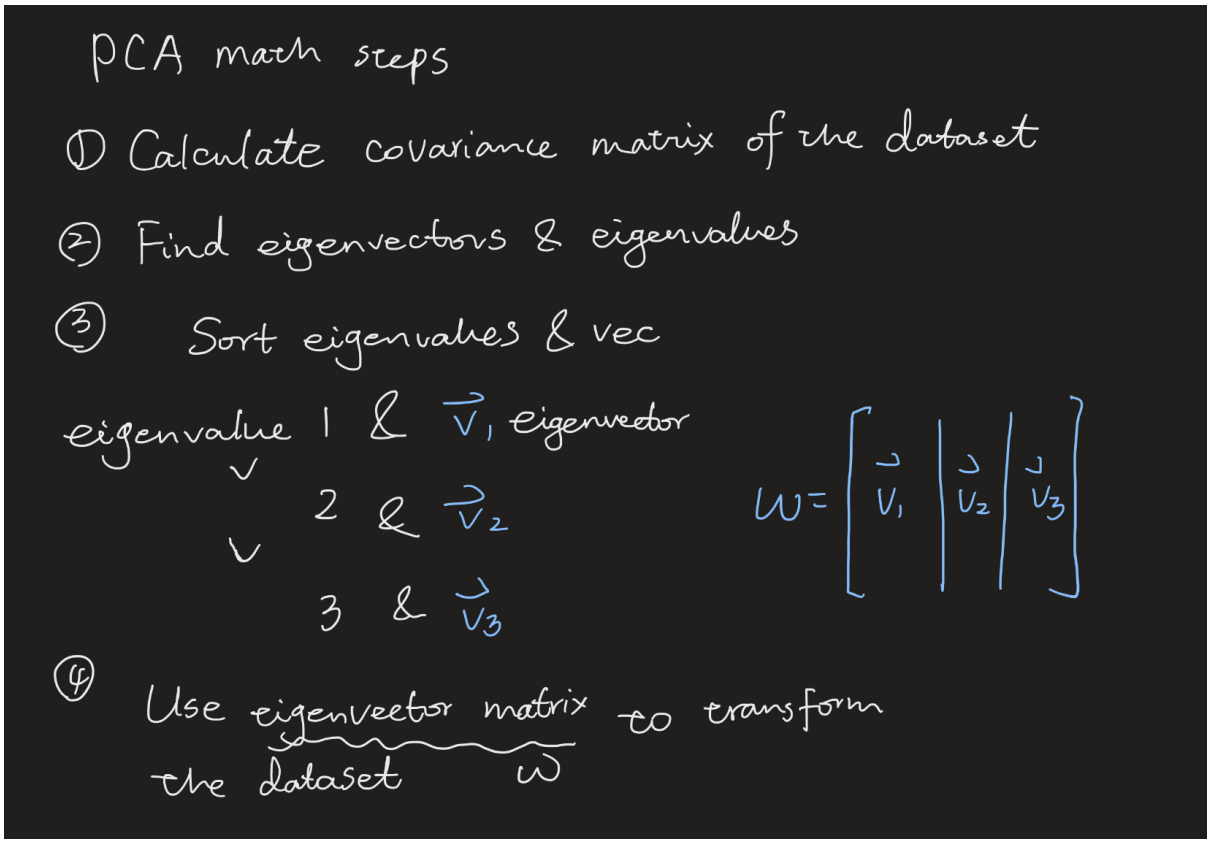
▼ **Preclass: prove eigenvectors are orthogonal to each other**



https://stats.stackexchange.com/questions/266652/why-are-pca-eigenvectors-orthogonal-and-what-is-the-relation-to-the-pca-scores-b

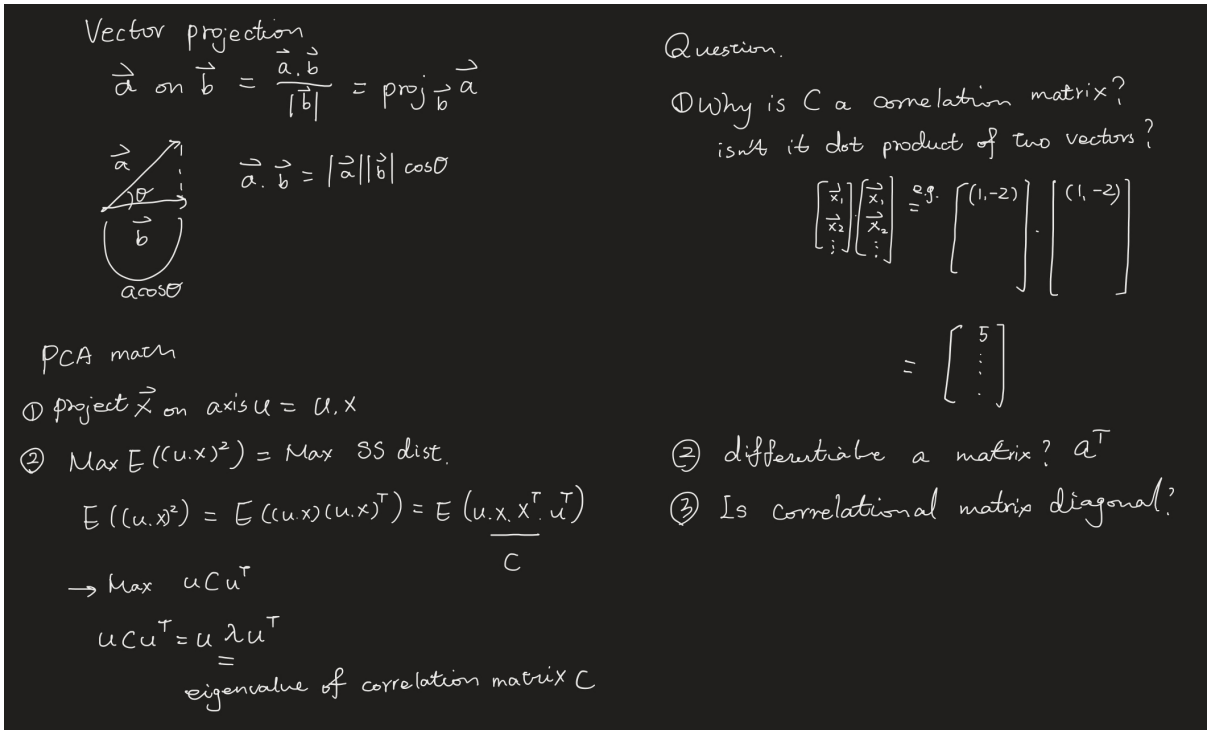
- why the biggest lambda will max the y?
- why we want cov matrix = 0 ?
  - bc cov matrix will be symmetric $x^T y$ = 0
- who is the eigen vector here? (y or a?)
- Why is the goal looks like that?

▼ Math steps to find PCA dimensions(aka eigenvector)



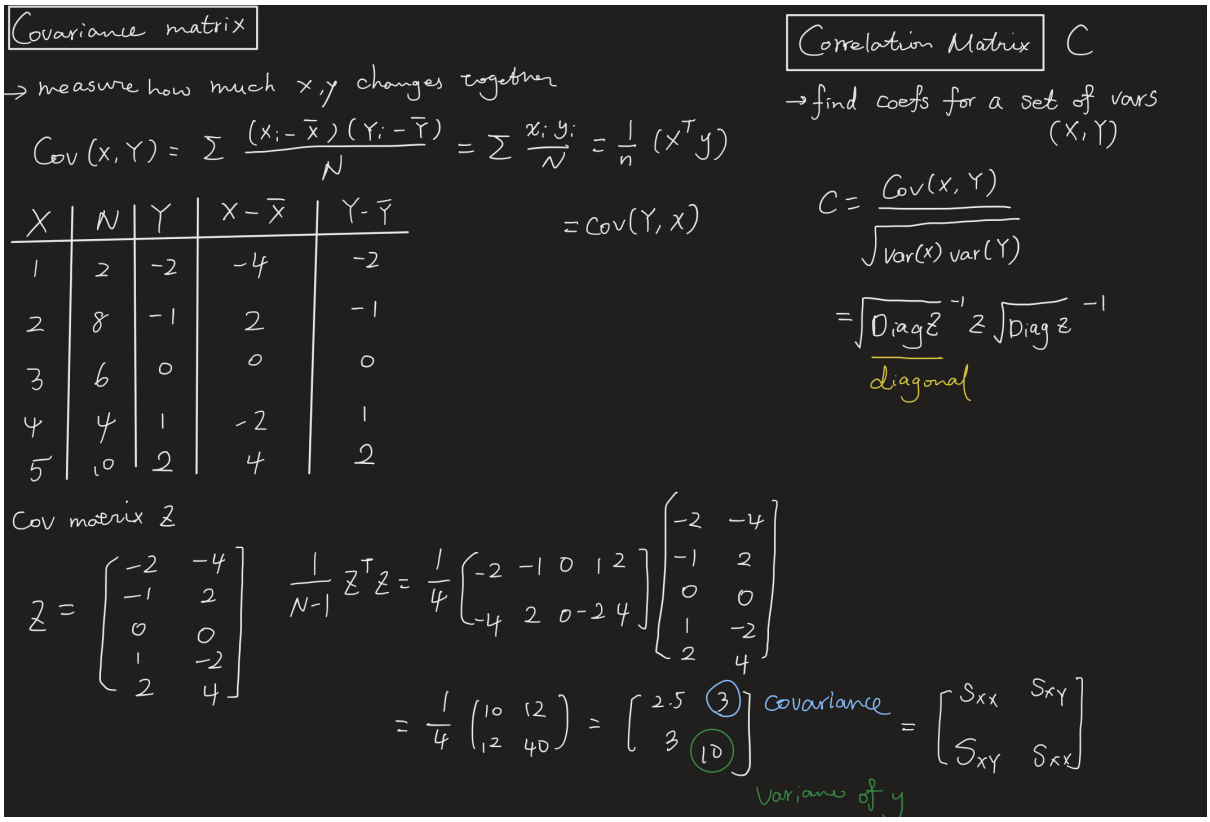
▼ PCA math foundations (Questions)

PCA math and question list

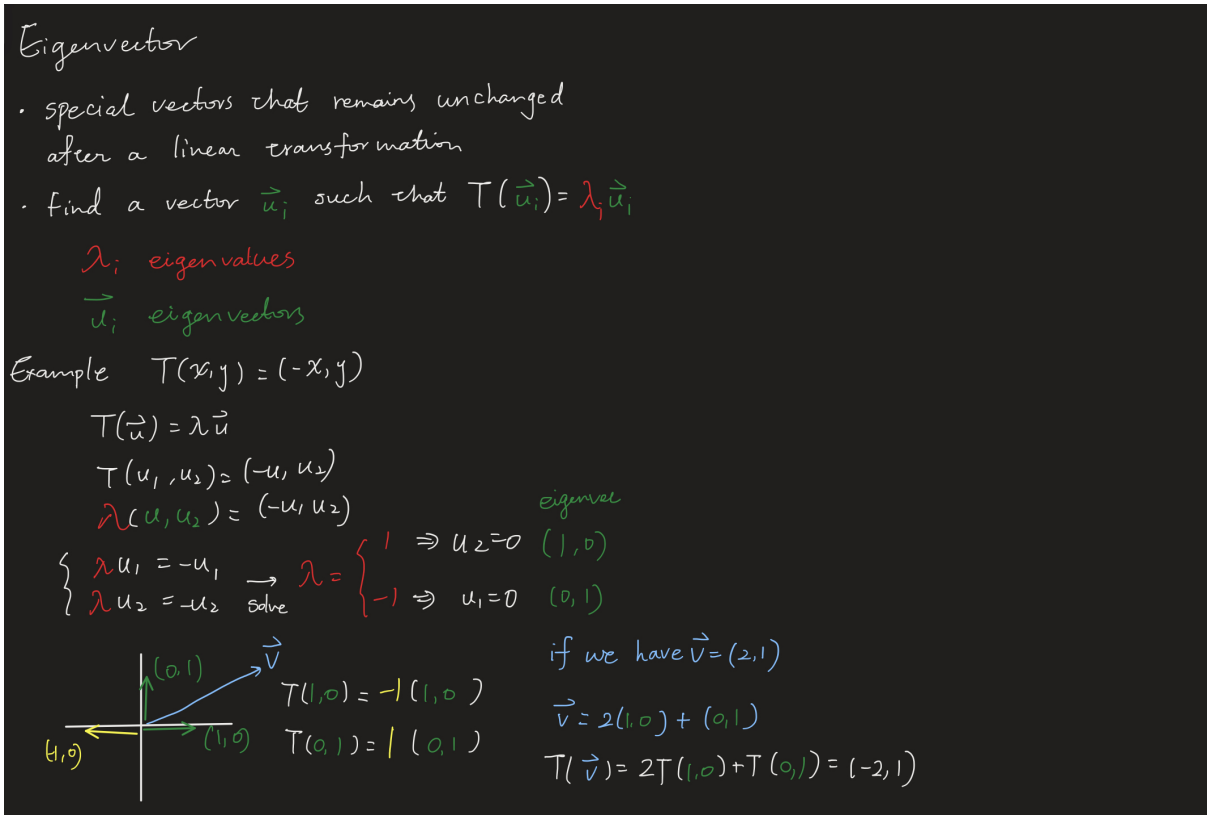


Connect to proof of eigenvectors are orthogonal to each other.

Covariance matrix & Correlation matrix



Eigenvector review



Math about coef & cov & eigen

The Mathematics Behind Principal Component Analysis

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible of the

https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643

coef matrix

How to express a correlation matrix in terms of a covariance matrix?

I'd like to know, if I have the following Normal multivariate structure $$\left({\begin{array}{*{20}{c}} {{Y_{r_1}}}\\ {{Y_{r_2}}} \end{array}} \right)\sim{\mathcal{N}_{r}}\left({\left[{\begin{a...

https://stats.stackexchange.com/questions/413033/how-to-express-a-correlation-matrix-in-terms-of-a-covariance-matrix

E((u.x)^2)

https://www.math.uci.edu/icamp/courses/math77b/lecture_12w/pdfs/PCA.pdf

Finn's note: here

Pre-class

1. Barber Exercise

Exercise 15.2. *Consider a dataset in two dimensions where the data lies on the circumference of a circle of unit radius. What would be the effect of using PCA on this dataset, in which we attempt to reduce the dimensionality to 1? Suggest an alternative one dimensional representation of the data.*

If the data is evenly distributed, as long as the PCA dimension crosses the origin, the effect of PCA will all be similar. Another method for dimensional representation is to use radio basis function to transform circular data into a straight line.