Interpolation/Extrapolation:

Add data points within two extremas and beyond without breaking the original trends.

Data Smoothing:

Moving Average Filtering:

Each point in a signal calculates the average of its neighboring points to give a new point.
Example: 3-point averaging (i.e. window size = 3)

$$x(avg) = \frac{x_{i-1} + x_i + x_{i+1}}{3}$$

HW: Looks for (i) centered moving average, (ii) weighted moving average, (iii) convolution.

Savitzky-Golay Filtering:

Smoothed data point:

$$(y_k)_s = \frac{\sum_{i=-n}^{n} A_i y_{k+i}}{\sum_{i=-n}^{n} A_i}$$

where A_{-n}, $A_{-(n-1)}$ …, A_{n-1}, A_n are weighting coefficients, known as convolution integers.

Ensemble Average:

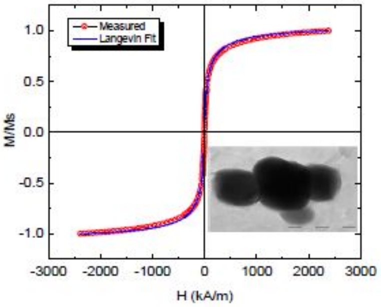
Average of N scans rather than averaging over a defined window around a single data point within a measurement.
Its possible to reduce the noise can be reduced by $\sqrt{N}$ by averaging N full scans that ultimately improves the signal to noise ratio (SNR).

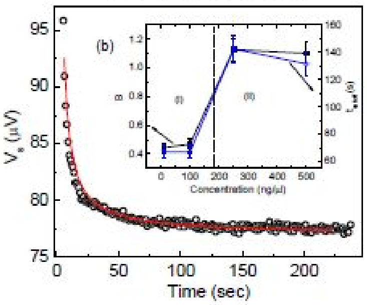
FFT Smoothing Filter:

Data smooth, Time-Frequency Conversion

Data Fitting:

Data can be visualized in a trend such as Linear, polynomial, exponential etc. and coefficients of the fit can be extracted so that the information will be useful for future experiment.
(Ref: *Devkota et. al., RSC Adv.*, 2015,5, 51169-51175)





Principal Component Analysis (PCA):

A dimension reduction technique while preserving much information contained in the original data set.

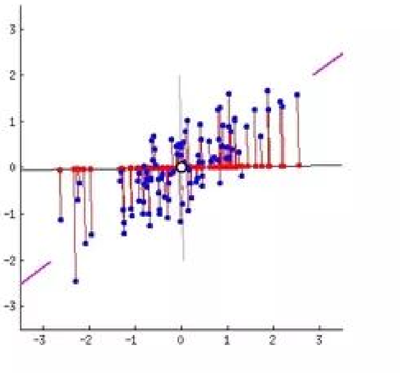
PCA considers linear transformation of observed data to identify desired new set of data for analysis.

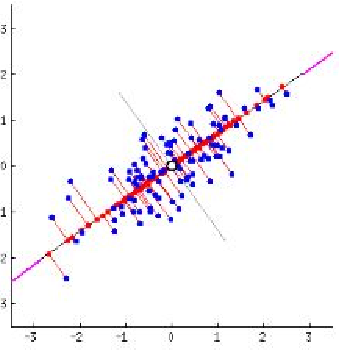
The transformed data has as large as possible variances in the first few dimensions so that these contain most information of the original data.

The most information is preserved when the variance of the projected data is maximum.

Also, the projected errors are to be minimized whenever possible!

PCA considers finding appropriate projected direction for the maximum variance and minimum error in projected data.





In the second choice:

Newly projected red points are more widely spread out than the first case. i.e. more variance

The projection error is less than that in the first case.

So, the second straight line (direction or axis) is favored.

General Mathematical Concept:

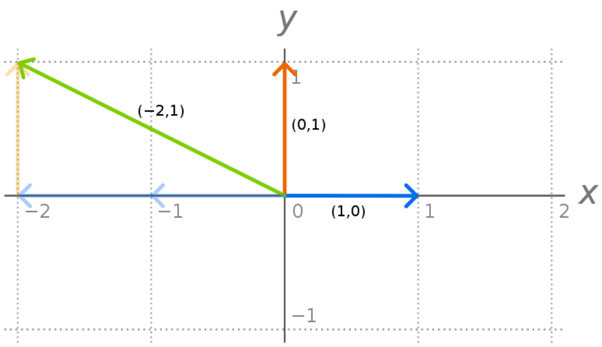
PCA uses linear transformation of matrices (covariance/scattering) to determine the new data set.

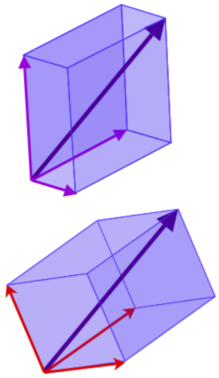
Linear Transformation of Matrix: AX = Y

Matrix A(2,3)*X(3,1) = Y(2,1): Dimension changes!

Eigenvalue and Eigenvectors.

Basis Vectors:

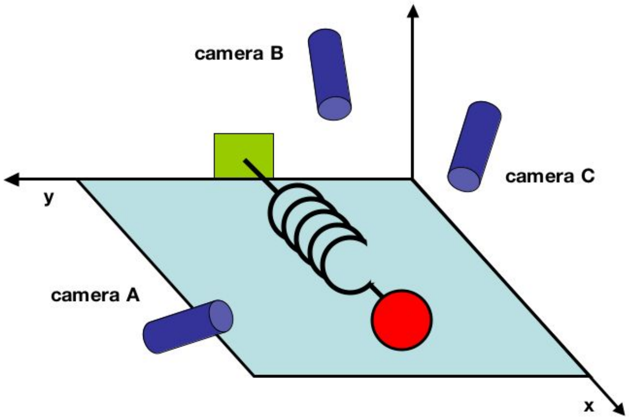




Change of Basis Vectors of a Vector in Vector Space: If vector R in vector space V can be expressed in terms of two basis Bold and Bnew, then it is possible to express the coordinate of the old basis with respect to the new basis.
Say, X and Y represent the basis vectors in Bnew and Bold, then Y=AX, where A is the transformation matrix.

References:
https://www.cc.gatech.edu/~lsong/teaching/CX4240spring16/pca_schlens.pdf
https://medium.com/district-data-labs/principal-component-analysis-with-python-4962cd026465
https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376

Toy example: Spring Pendulum
We know spring pendulum (with a ball of mass **M** and frictionless, massless spring) oscillates in one direction only. In Fig below, along x-direction.
However, if a person who does not know anything about the ‘physics’ of a pendulum, wants to record its displacement in x,y direction using three cameras as shown.



Spring oscillator

$$\vec{X} = \begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \\ x_C \\ y_C \end{bmatrix}$$

Vector representation for one trial (each camara produces 2D projection of the ball)

Each X represents total measurement types, say ***m***. What about for ***n*** trial?: Gets ***n*** column vectors.

Noise and Rotation:

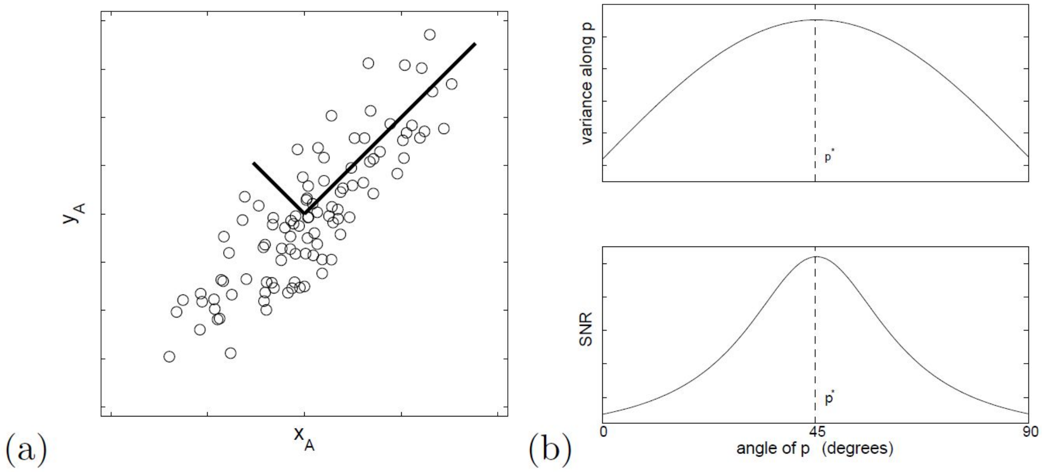
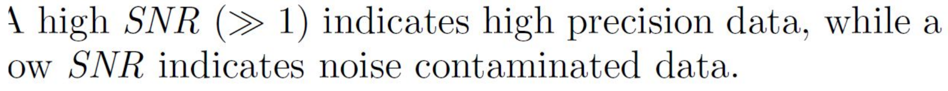


FIG. 2 (a) Simulated data of (x_A, y_A) for camera A. The signal and noise variances σ^2_{signal} and σ^2_{noise} are graphically represented by the two lines subtending the cloud of data. (b) Rotating these axes finds an optimal p^* where the variance and SNR are maximized. The SNR is defined as the ratio of the variance along p^* and the variance in the perpindicular direction.

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}. \qquad (2)$$


\ high SNR ($\gg 1$) indicates high precision data, while a
ow SNR indicates noise contaminated data.

In Fig. 2(a), the signal is along the longer line and the noise is along the short line - variance of the points along the longer line is maximum and that along the short line is minimum. So, the dimension represented by the longer line is the important dimension in this example - a principal component not to omit.

In a multi-dimensional problem, one first finds the axis associated with the maximum variance, then the second axis with the second maximum variance and perpendicular to the first. jth axis will be such that it will have jth maximum variance and will be perpendicular to all previous axes. At the end, one can ignore most axes and consider which contain the most information.

Why is omitting a few dimensions not a problem?

1. Few dimensions can be expressed in terms of others if there are correlations.
2. Most information is captured by the first few dimensions that have maximum variance.

Redundancy and Dimension Reduction:

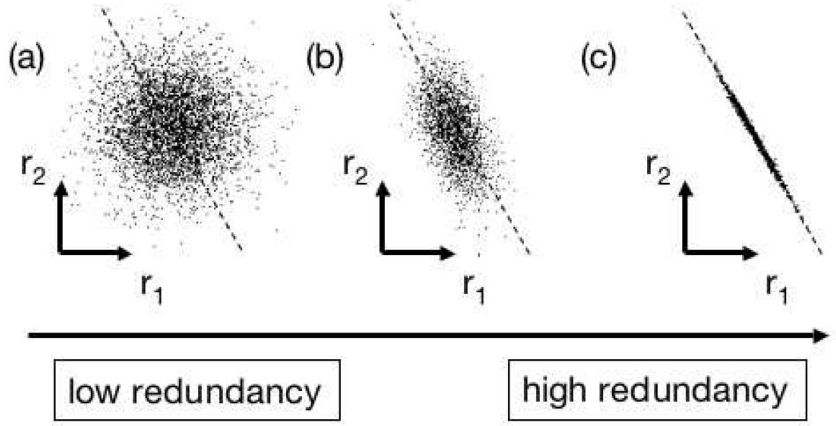
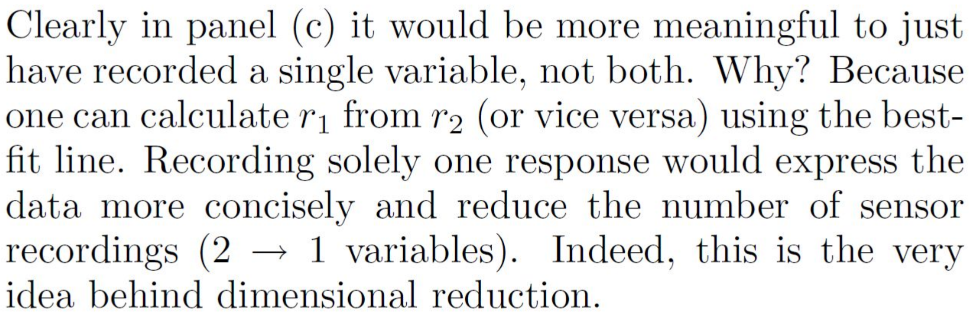


FIG. 3 A spectrum of possible redundancies in data from the two separate recordings r_1 and r_2 (e.g. x_A, y_B). The best-fit line $r_2 = kr_1$ is indicated by the dashed line.



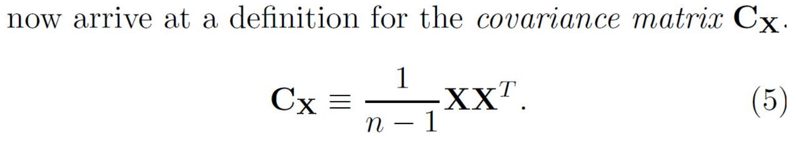
Clearly in panel (c) it would be more meaningful to just have recorded a single variable, not both. Why? Because one can calculate r_1 from r_2 (or vice versa) using the best-fit line. Recording solely one response would express the data more concisely and reduce the number of sensor recordings ($2 \rightarrow 1$ variables). Indeed, this is the very idea behind dimensional reduction.

STEPS:

1. Standardize the data.

   (If not, differences in the measurement length or other variation create problems - basically this means normalizing the data to get variance in a fixed range for all data sets.)

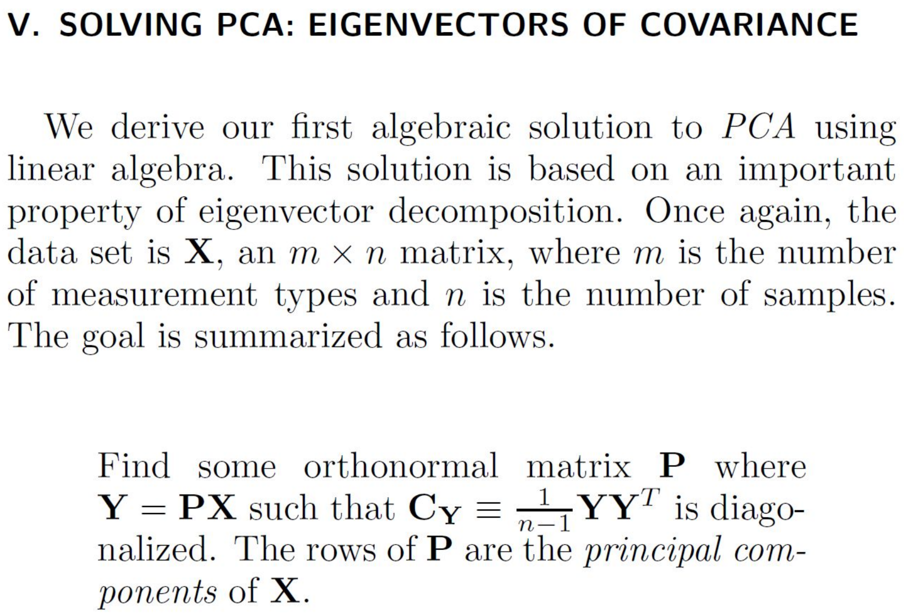
2. Use the standardized data to generate a covariance matrix (or perform Singular Vector Decomposition).



now arrive at a definition for the *covariance matrix* $\mathbf{C_X}$.

$$\mathbf{C_X} \equiv \frac{1}{n-1}\mathbf{X}\mathbf{X}^T. \qquad (5)$$

3. Obtain eigenvectors (principal components) and eigenvalues from the covariance matrix. Each eigenvector will have a corresponding eigenvalue.



V. SOLVING PCA: EIGENVECTORS OF COVARIANCE

We derive our first algebraic solution to *PCA* using linear algebra. This solution is based on an important property of eigenvector decomposition. Once again, the data set is $\mathbf{X}$, an $m \times n$ matrix, where m is the number of measurement types and n is the number of samples. The goal is summarized as follows.

> Find some orthonormal matrix $\mathbf{P}$ where $\mathbf{Y} = \mathbf{P}\mathbf{X}$ such that $\mathbf{C_Y} \equiv \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T$ is diagonalized. The rows of $\mathbf{P}$ are the *principal components* of $\mathbf{X}$.

4. Sort the eigenvalues in descending order.
5. Select the k eigenvectors with the largest eigenvalues, where k is the number of dimensions used in the new feature space (k≤d).
6. Construct a new matrix with the selected k eigenvectors.