ML in Healthcare

#C15 Principal component analysis (PCA)

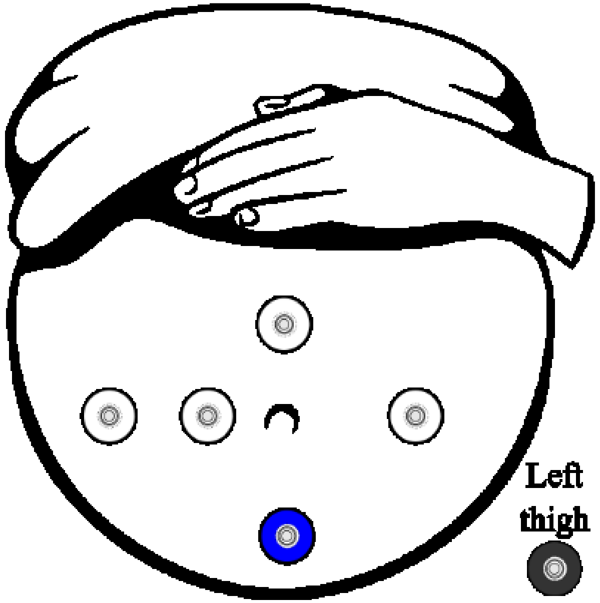
Technion-IIT, Haifa, Israel

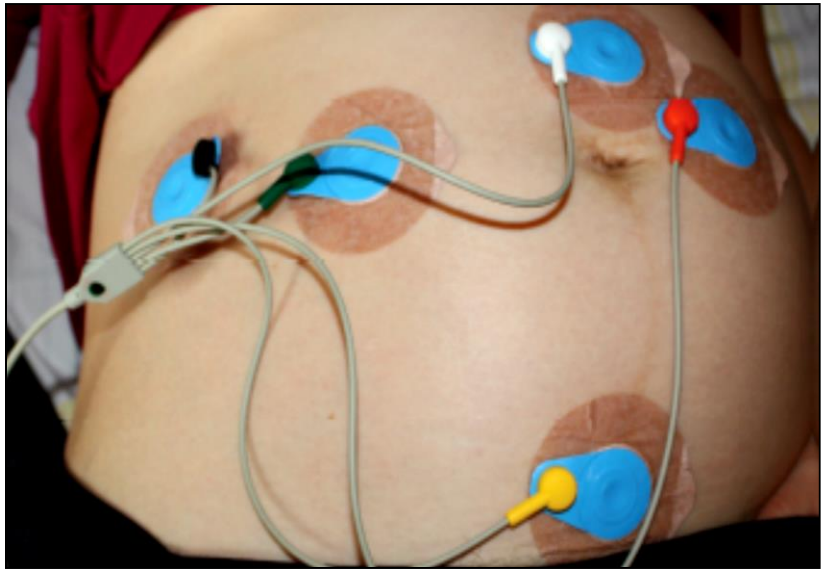
Assist. Prof. Joachim Behar
Biomedical Engineering Faculty
Technion-IIT

Topics covered (2 lectures)

- Blind Source Separation.
- Principal Component Analysis (PCA).
- PCA in Machine Learning.
- Independent Component Analysis (ICA).

NI-FECG



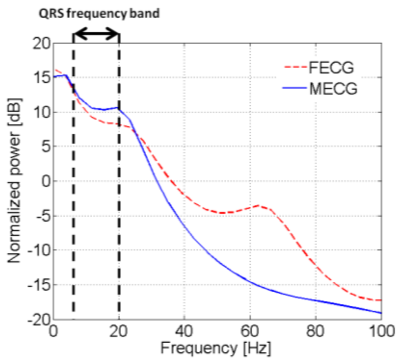


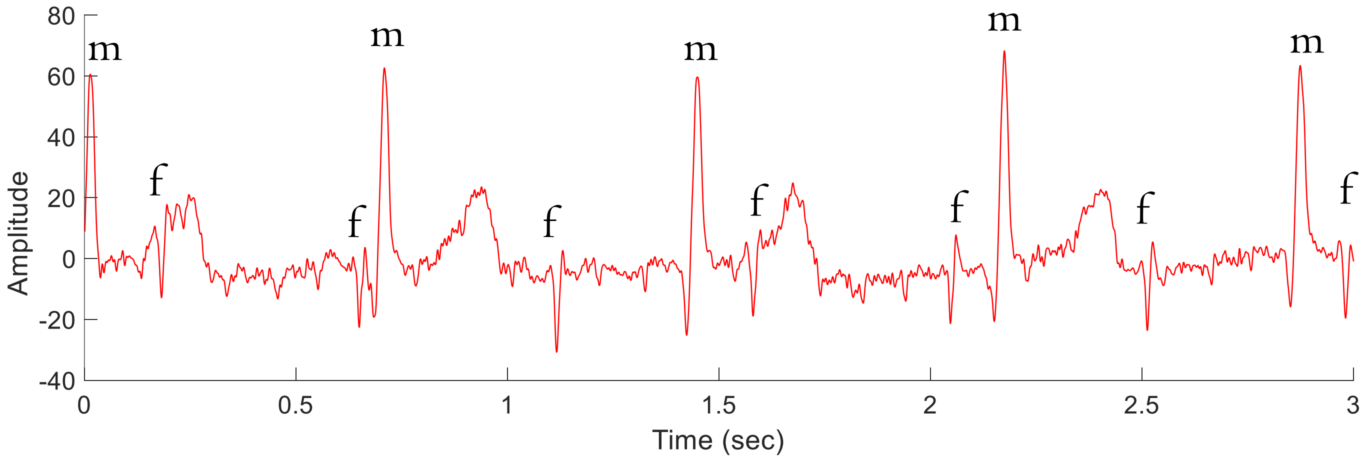
NI-FECG: opportunity

- Non-invasive,
- Information on conduction,
- Low-cost,
- Remote monitoring.

NI-FECG: Challenges

- Overlap in time and frequency,
- Non stationarities,
- Vernix caseosa.





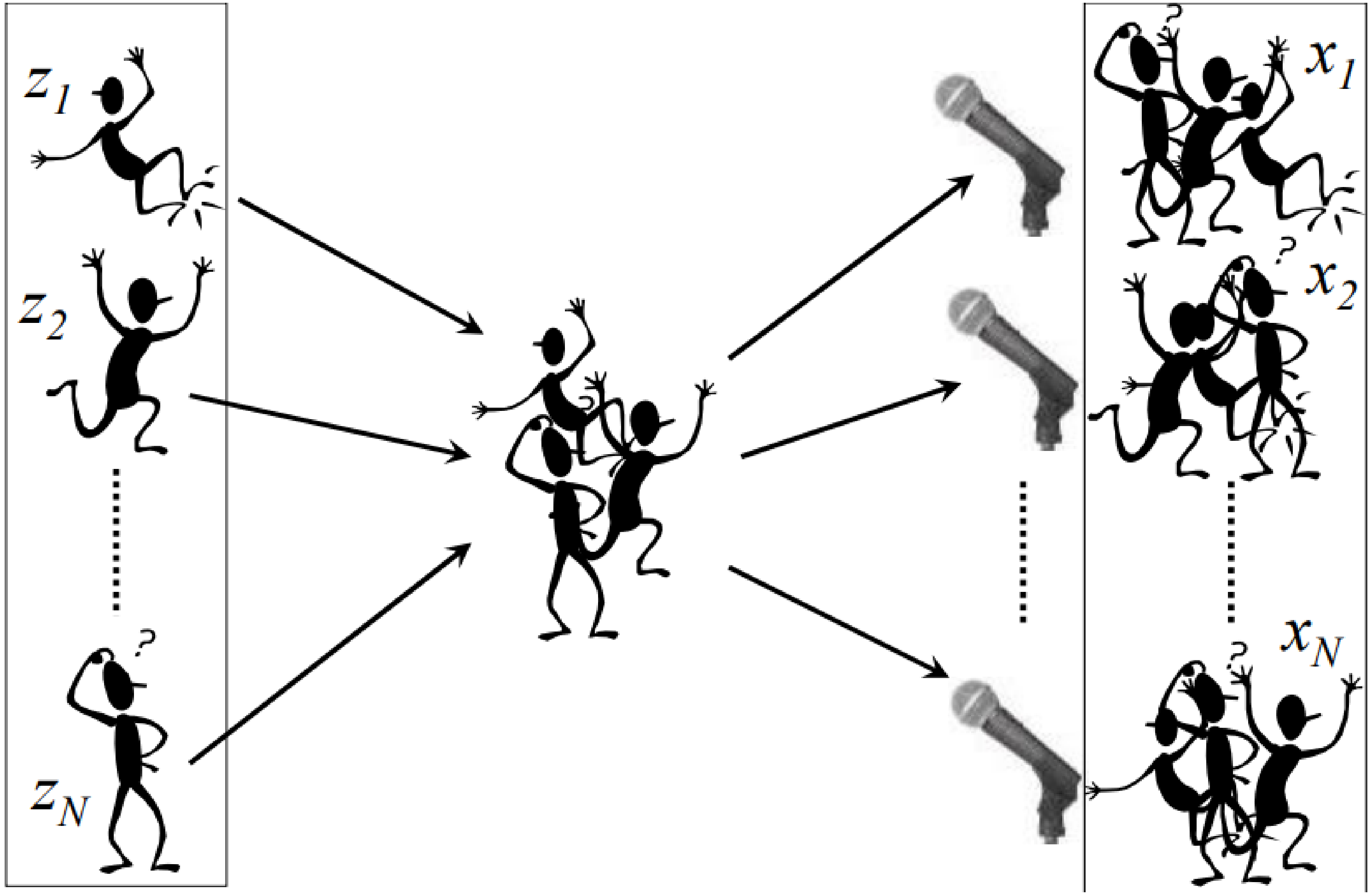
FECGSYN

The effects of asymmetric volume conductor modeling on non-invasive fetal ECG extraction

Keenan, E., et. al. Physiological measurement, 39.10 2018
Parent models: Behar et al 2014 and Sameni et al. 2007.

Blind Source Separation

What is Blind Source Separation (BSS)?

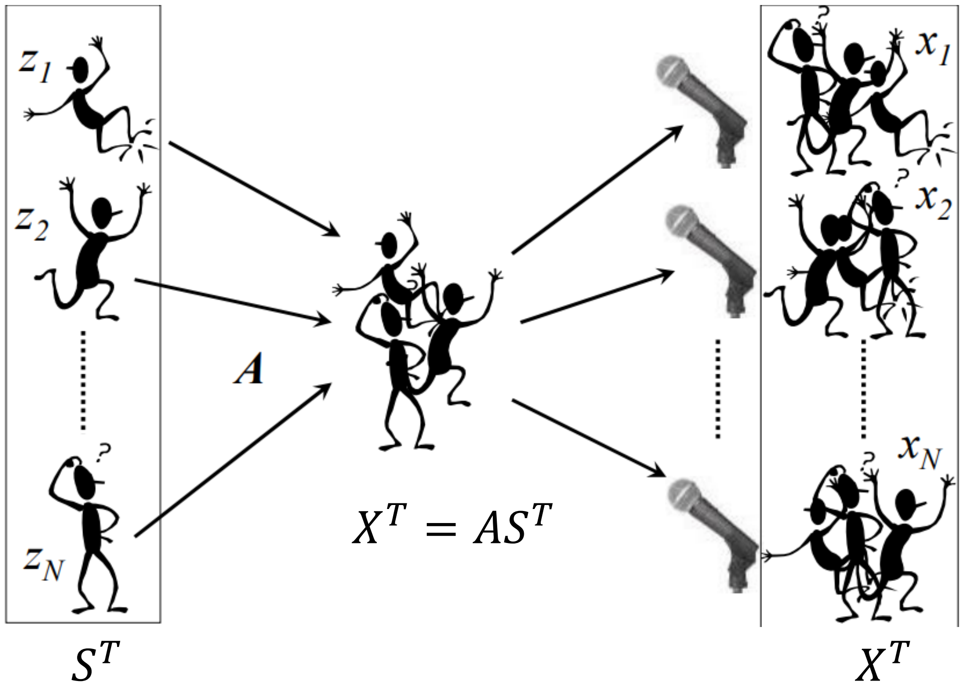


Example: https://cnl.salk.edu/~tewon/Blind/blind_audio.html

What is Blind Source Separation (BSS)?

- Assume a set of observed signals which are **linear mixture** of unknown **independent** source signals.
- The mixing (not the signals) is stationary.
- We have as many observed signals as unknown sources.
- BSS aims to recover the original **independent** sources from the observed **linear** mixtures.
- In other words, BSS consist of the evaluation of a set of source signals from a set of mixed signals. This is done without information (or very little) about the source signals or the way the signals are mixed together (mixing process).

The Cocktail party problem



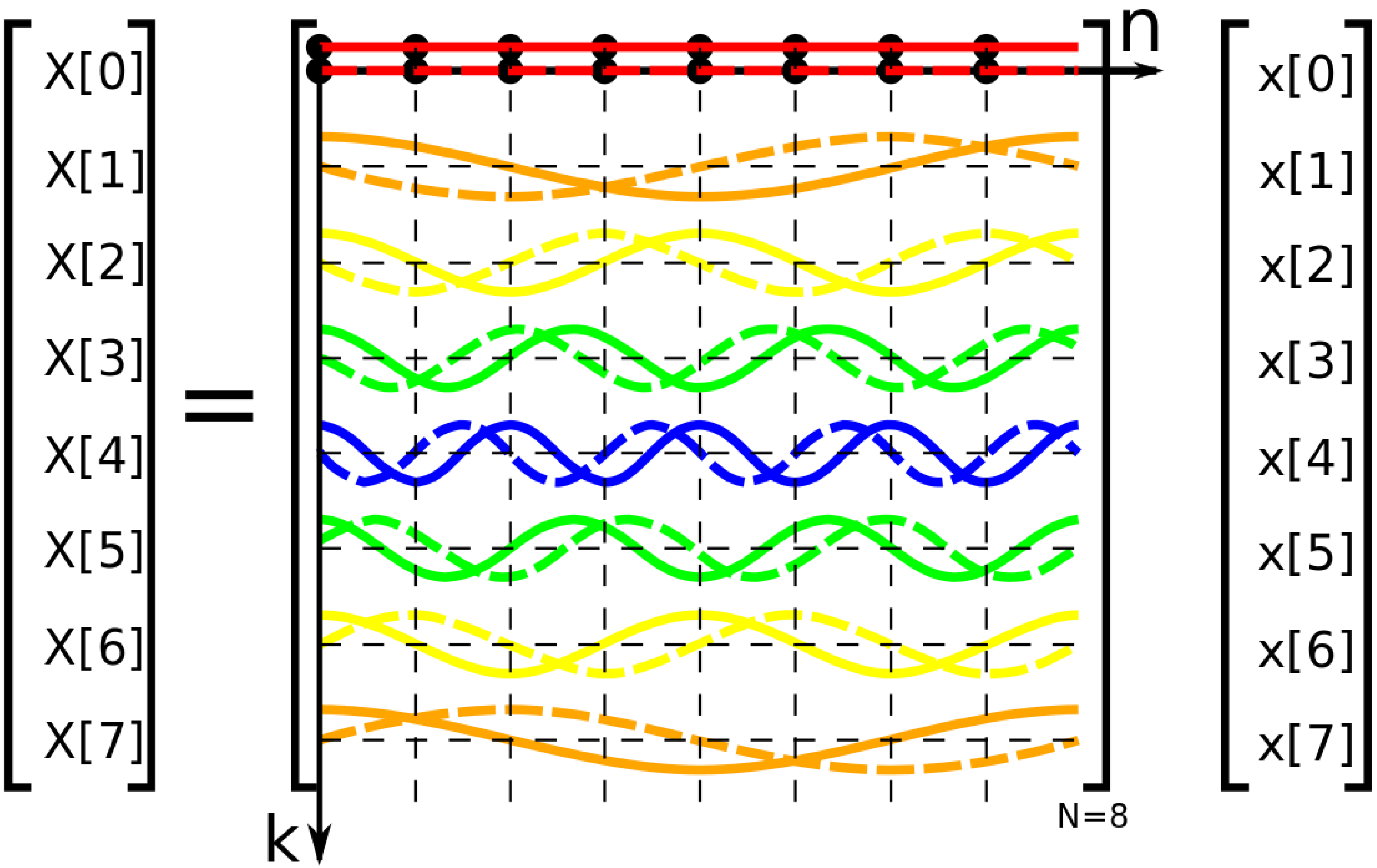
- At each time instant:
  - $x(t) = As(t)$ and $s(t) = Wx(t)$
- For all recorded observations:
  - $X^T = AS^T$
  - $\hat{S}^T = WX^T$ with $W = \hat{A}^{-1}$
    - $A \in \mathbb{R}^{n \cdot n}$: linear square mixing.
    - $X \in \mathbb{R}^{m \cdot n}$: observations produced by the mixing.
    - $S \in \mathbb{R}^{m \cdot n}$: independent sources.
    - n sources and observed signals.
    - m observations (datapoint).
- We want to estimate $W = \hat{A}^{-1}$.
- For that purpose we need a way to **measure independence**.

Analogy with Fourier transform

- Fast Fourier Transform (FFT):
  - $S = Wx$
  - x: the original input signal (analogy sensor in BSS).
  - $W \in \mathbb{R}^{n \cdot n}$: square DFT matrix.
  - S: the DFT of the signal (analogy source in BSS).
- Square DFT matrix expansion, $\omega = e^{-2\pi i/N}$

$$W = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(N-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots & \omega^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \omega^{3(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix}$$

Analogy with Fourier transform

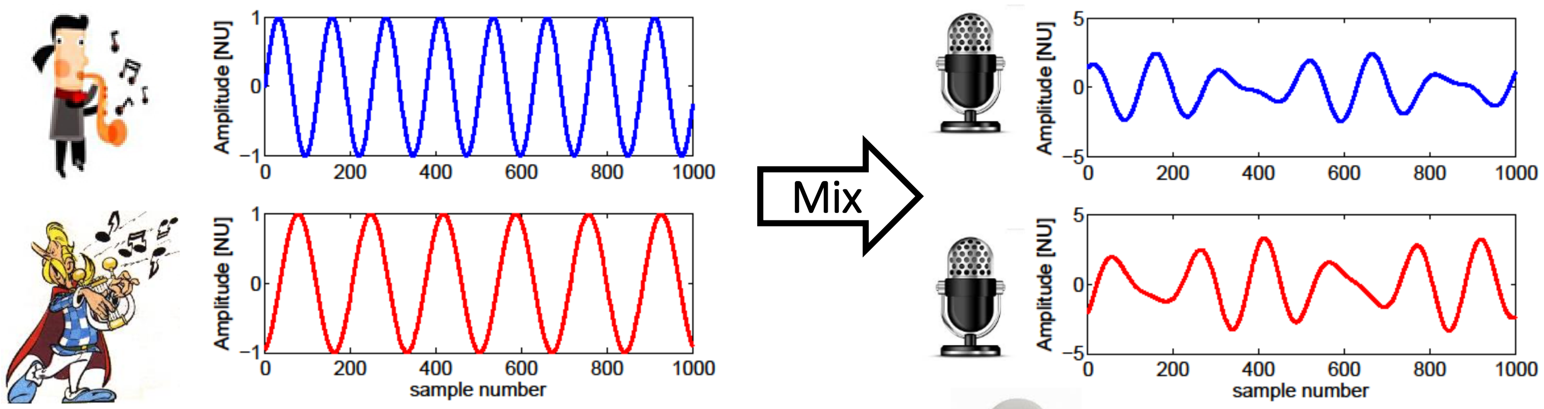


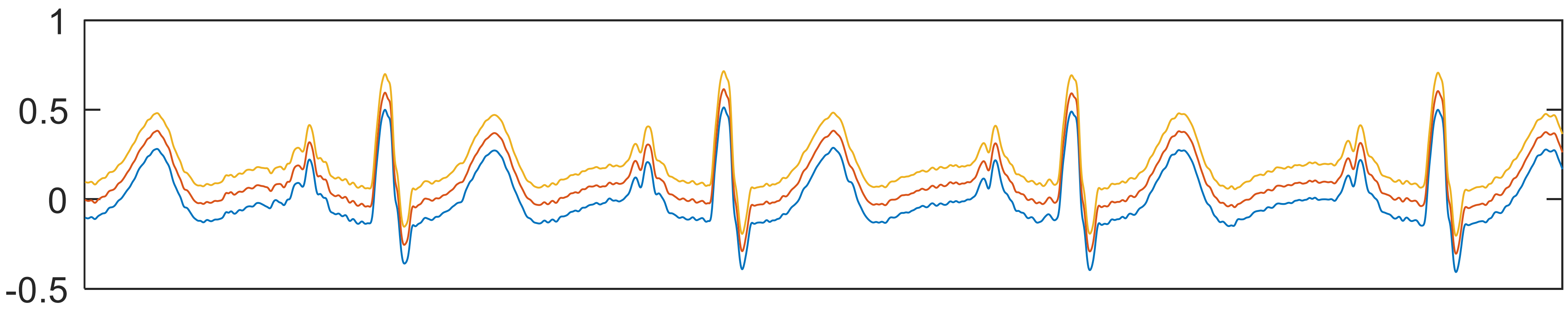
https://commons.wikimedia.org/wiki/File:Fourierop_rows_only.svg

Analogy with Fourier transform

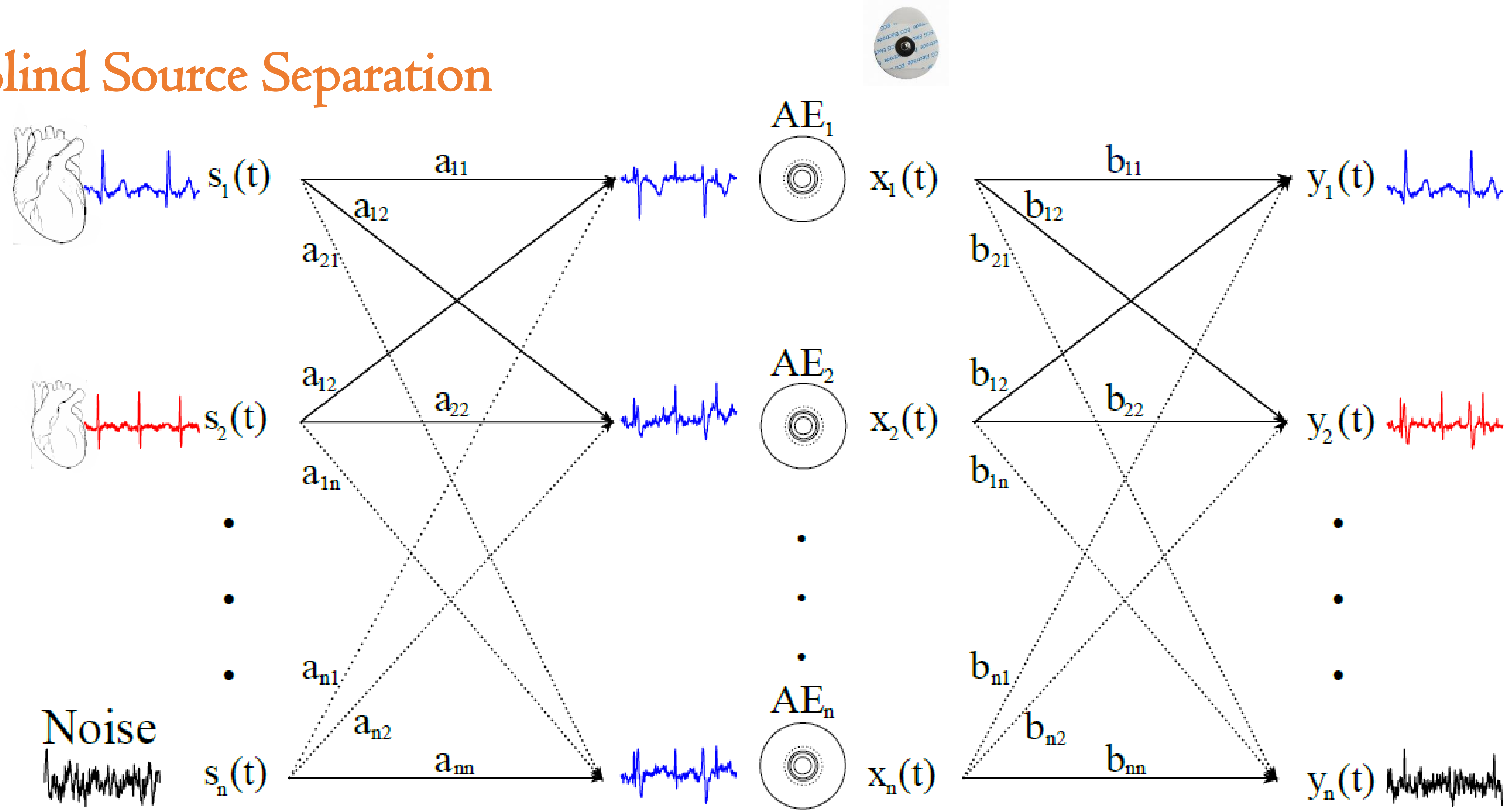
- **Like** FFT, with BSS, we decompose our sensed signals by transforming the observations into another vector space which maximize the separation between interesting signal and unwanted noise.
- **Unlike** FFT, this separation is not based on frequency but independence.
- In BSS we only assume **independence and linear mixing**.
- So sources may have the same frequency content and one can filter/separate in-band noise/signals with BSS.
- We will study two widely used techniques for BSS:
  - Principal Component Analysis (PCA).
  - Independent Component Analysis (ICA).
- We will use the maternal-fetal ECG mixing as our toy example.

Blind Source Separation





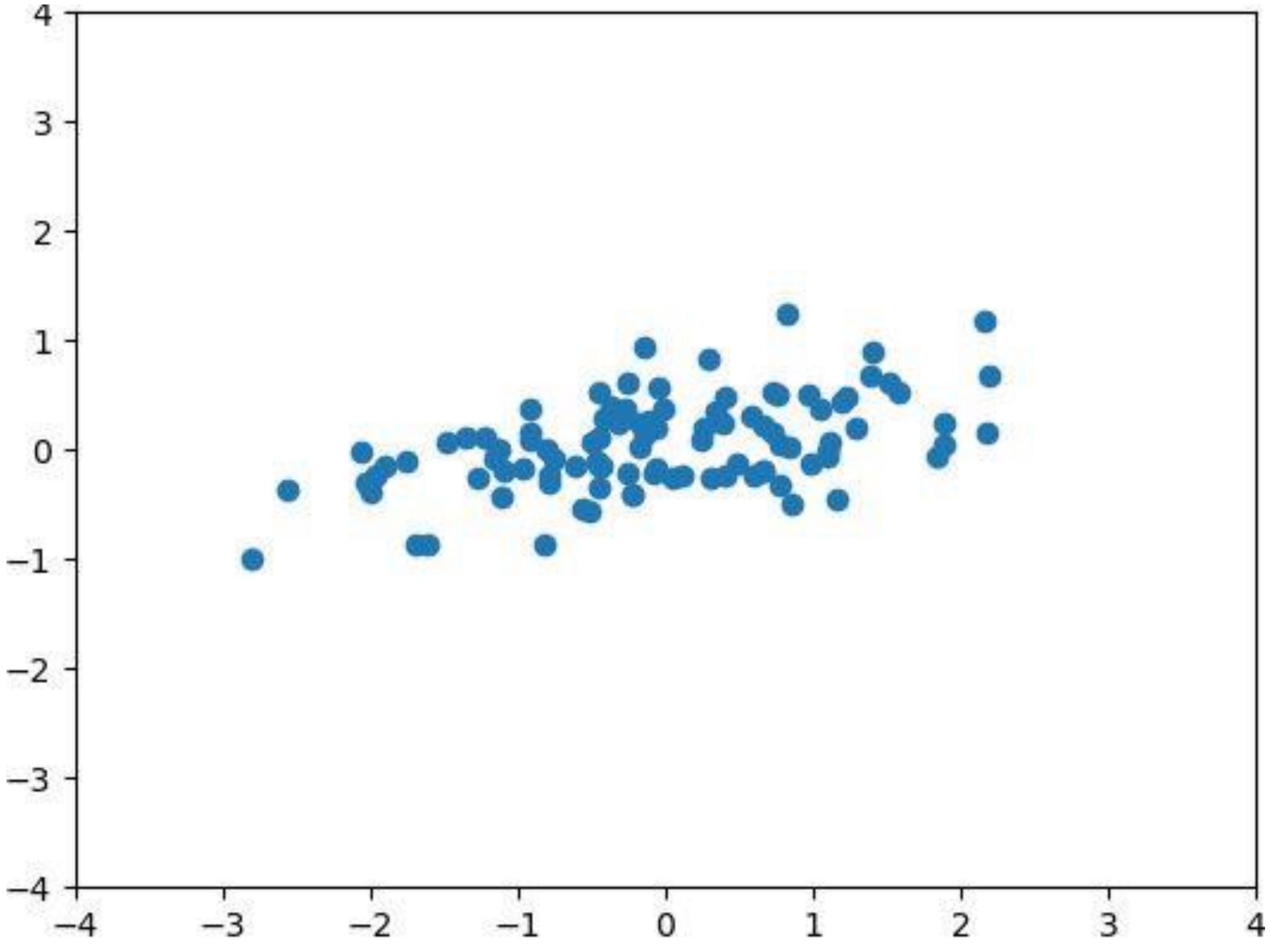
Blind Source Separation



AIMLab.

Change of basis

- Is there a better way to represent the data?



https://medium.com/x8-the-ai-community/principal-component-analysis-a-brief-introduction-dc8cf3e03c71

Principal component analysis

Common definition of PCA: Minimum Error Formulation

Codename
Karl Pearson
1857-1936

Special power
Statistics Master

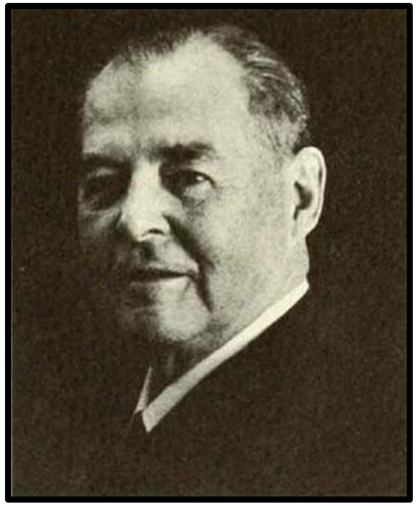
Place of origin

PCA

Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space"(PDF). *Philosophical Magazine* 2 (11): 559–572.

Definition: The linear projection that minimizes the average projection cost, defined as the mean squared distance between the data points and their projections.

Common definition of PCA: Maximum Variance Formulation



Codename

Harold Hoteling

1895-1973

Special power

Statistics Master

Place of origin

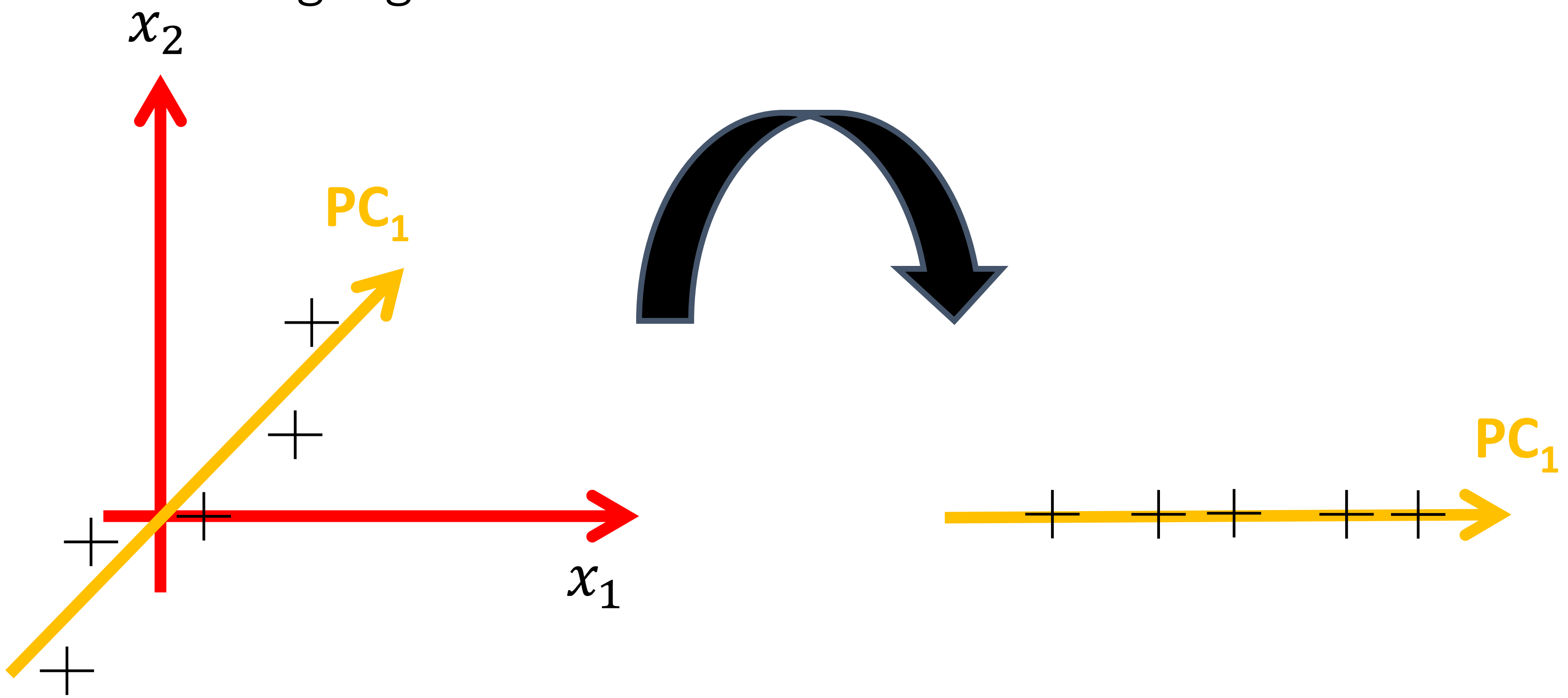
PCA

Hotelling, H (1936). "Relations between two sets of variates". Biometrika. 28 (3/4): 321–377.

Definition: The orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such as that the variance of the projected data is maximized.

Principal Component Analysis

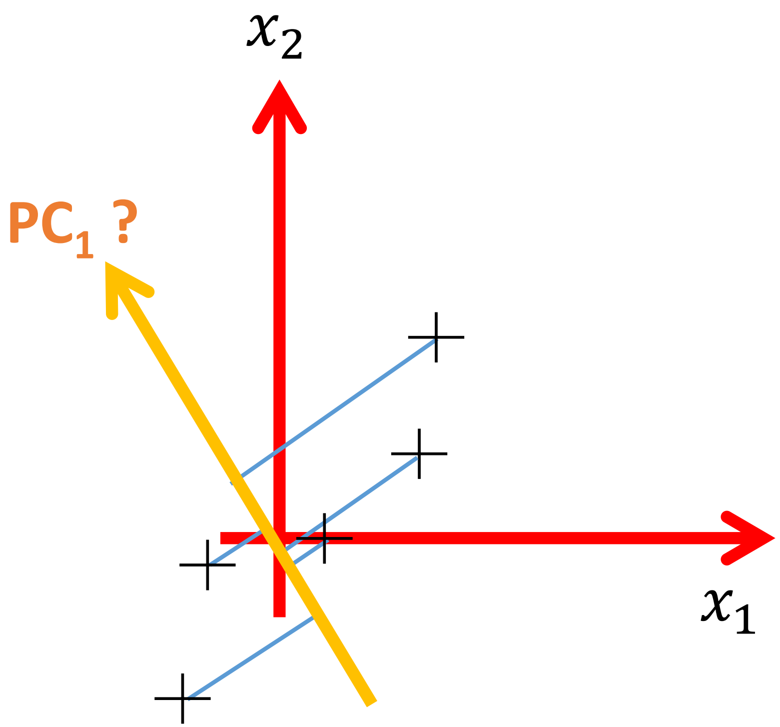
- To identify the most meaningful basis in some sense to re-express a data set. In this basis it is expected that hidden structure will be revealed or that the important structure will be better highlighted.

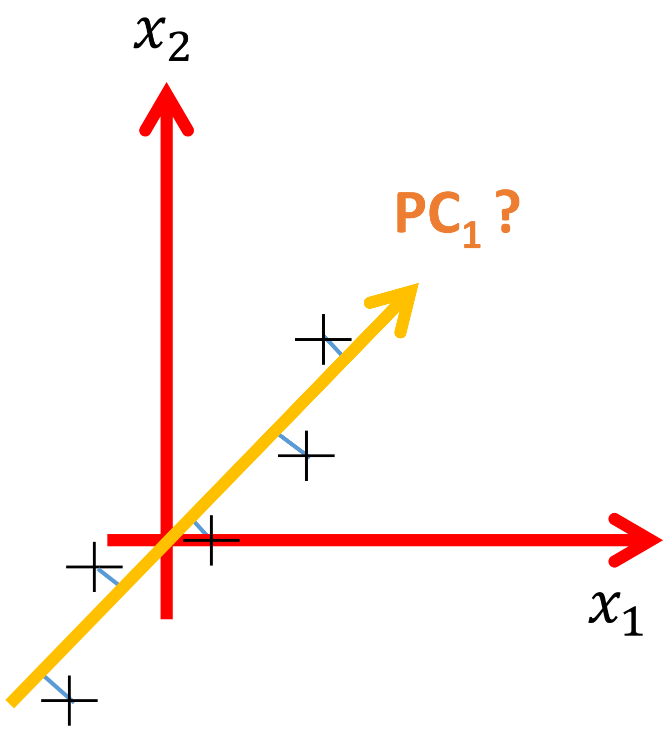


Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space"(PDF). Philosophical Magazine 2 (11): 559–572.
Hotelling, H (1936). "Relations between two sets of variates". Biometrika. 28 (3/4): 321–377.

Principal Component Analysis

- Hotelling definition: “The orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, **such as that the variance of the projected data is maximized**”.

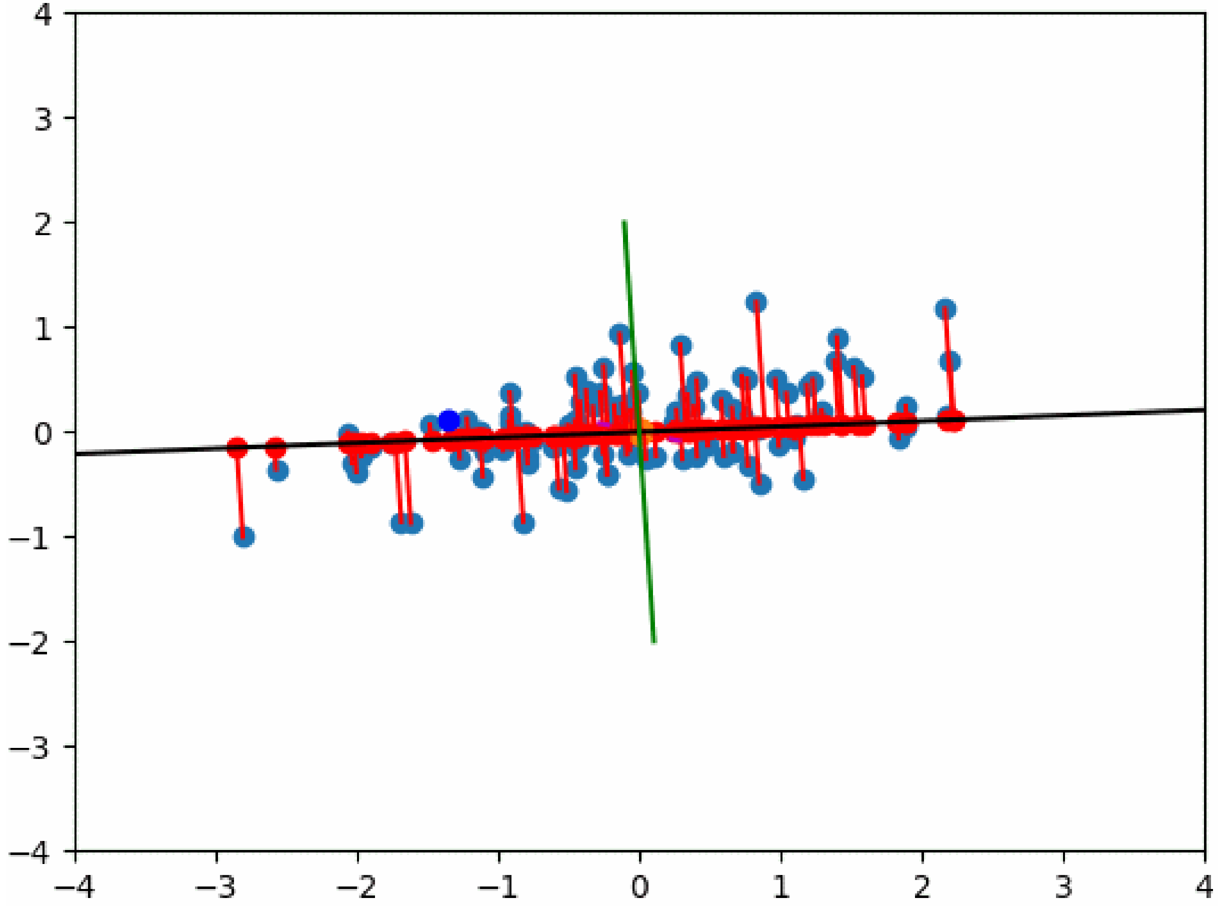




AIMLab.

Change of basis

- Is there a better way to represent the data?



https://medium.com/x8-the-ai-community/principal-component-analysis-a-brief-introduction-dc8cf3e03c71

Principal Component Analysis

- Math for PCA:
  - We assume a set of observation.
  - We look for the principal component $u_1 \in \mathbb{R}^d$ and assume a unit vector $u_1^T u_1 = 1$.
  - Datapoints sample set mean is: $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$.
  - The variance of the projected data: $\sigma_1^2 = \frac{1}{m}\sum_{i=1}^{m}[\,u_1^T x^{(i)} - u_1^T \bar{x}]^2 = u_1^T C u_1$.
    - Where C is the covariance matrix: $C = \frac{1}{m} X X^T$.
    - The covariance matrix generalize the notion of variance to multiple dimensions.
  - This is the quantity we want to maximize.
  - We now write our maximization problem as:
    - $max_{||u_1||=1}\{\,u_1^T C u_1\}$.
    - That is we look for maximizing the projected variance onto the new PC.
    - Under the constraint that $||u_1|| = 1$.

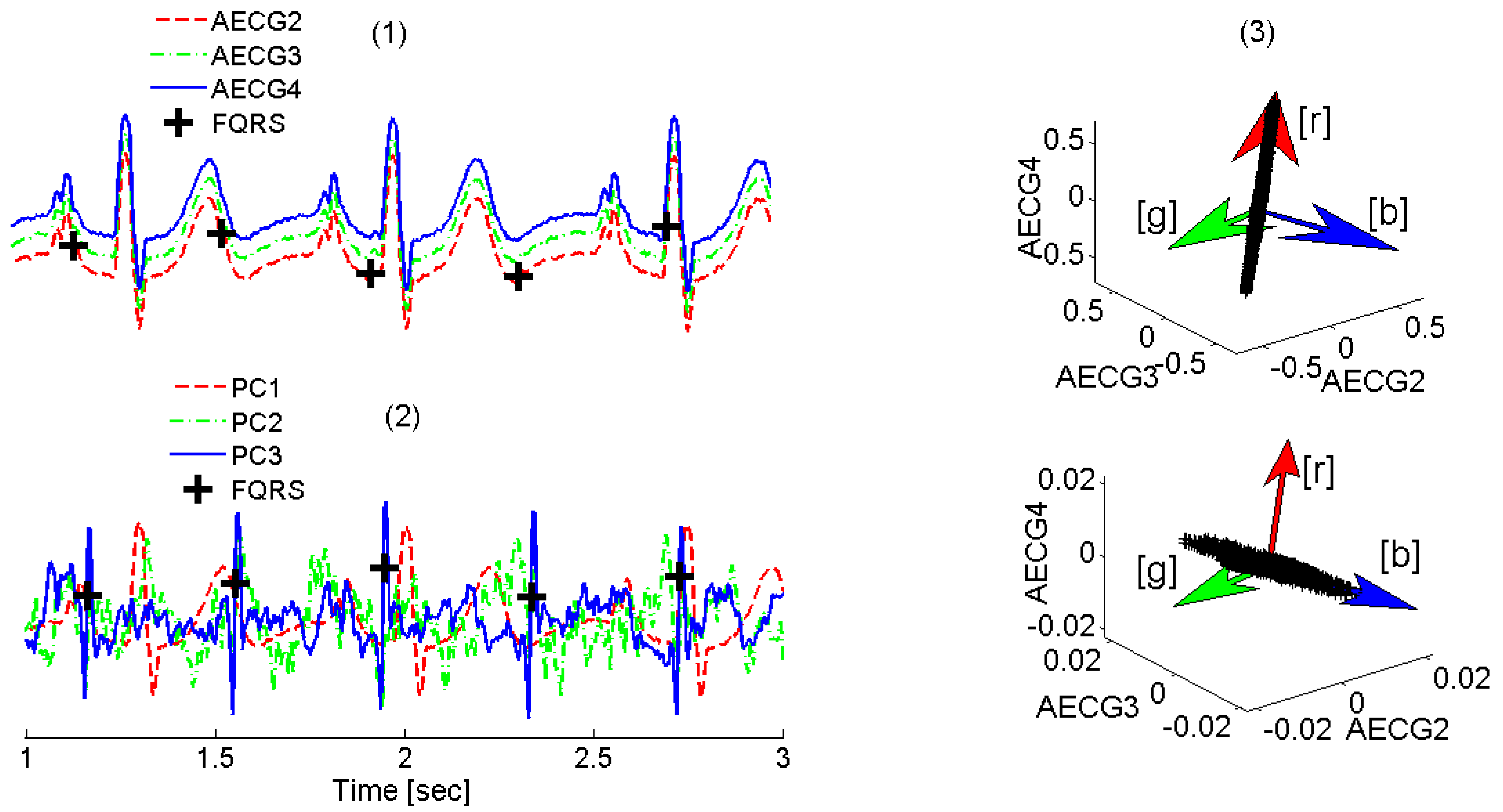
Principal Component Analysis

- Math for PCA:
  - For that purpose we use the Lagrange multiplier and make the unconstrained maximization of:
    - $L(u_1, \lambda) = u_1^T C u_1 + \lambda_1\left(1 - u_1^T u_1\right)$
    - $\frac{\partial L(u_1,\lambda)}{\partial u_1} = 0 \leftrightarrow Cu_1 = \lambda_1 u_1 \rightarrow u_1$ is an eigenvector of C with eigenvalue λ_1.
    - $u_1^T C u_1 = \lambda_1 \rightarrow$ Variance is maximal for the largest eigenvalue.

- We can define additional principal components in an incremental fashion by choosing each new direction to be that which maximize the projected variance amongst all possible directions orthogonal to those already considered.

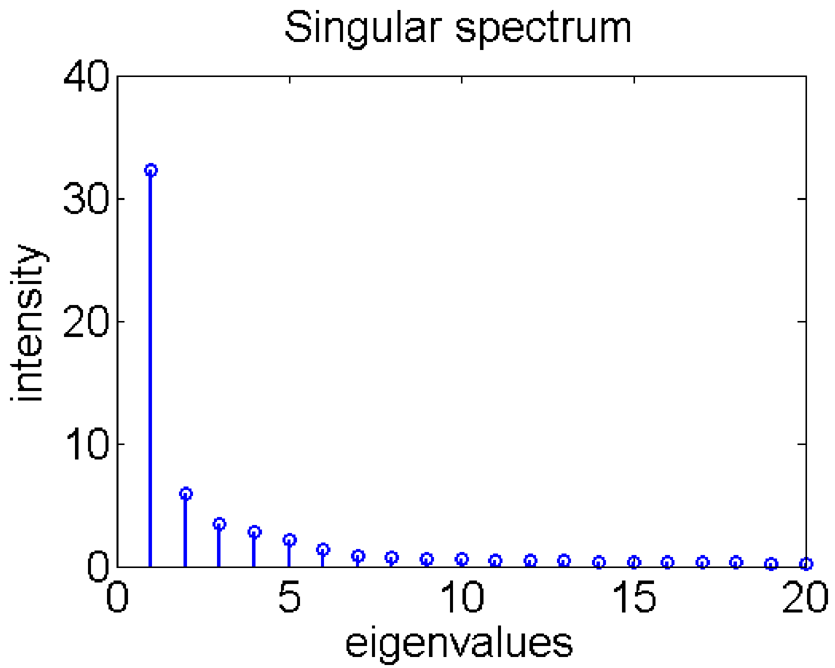
AIMLab.

Principal Component Analysis



Principal Component Analysis

- Eigenspectrum:
  - Magnitude of the projected data along each of the eigenvectors.
  - Recall: $u_1^T C u_1 = \lambda_1$ → **largest variance** = largest eigenvalue.
  - **Eigenspectrum** corresponds to the plot of the eigenvalues.
  - It provides a representation of how much "energy" (information) each eigenvector carry.



Principal Component Analysis

- Such a basis of eigenvectors will always exist and diagonalise C:
  - The covariance matrix C is a positive-semidefine and symmetric matrix.
  - A symmetric matrix can be diagonalised by an orthogonal matrix of its eigenvectors.
  - Thus using linear algebra:
    - $\exists P \in O(R, p) \ / \ C = PDP^T$ where D is diagonal.

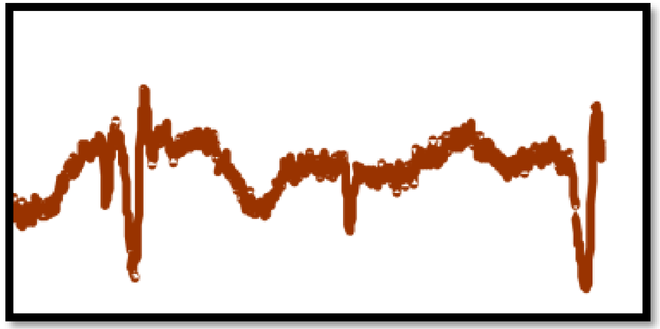
- Practically you can use **Singular Value Decomposition** (SVD) to find the PCA transform.

Principal Component Analysis

- In summary:
  - Eigenvectors of C: defines the PCA base.
  - Eigenvalues of C: define the order of importance of the PCAs.
  - Eigenvectors are orthogonal: the new basis is orthogonal.

- In practice, PCA can be used for:
  - For visualization,
  - Dimensionality reduction,
  - **Source separation**.

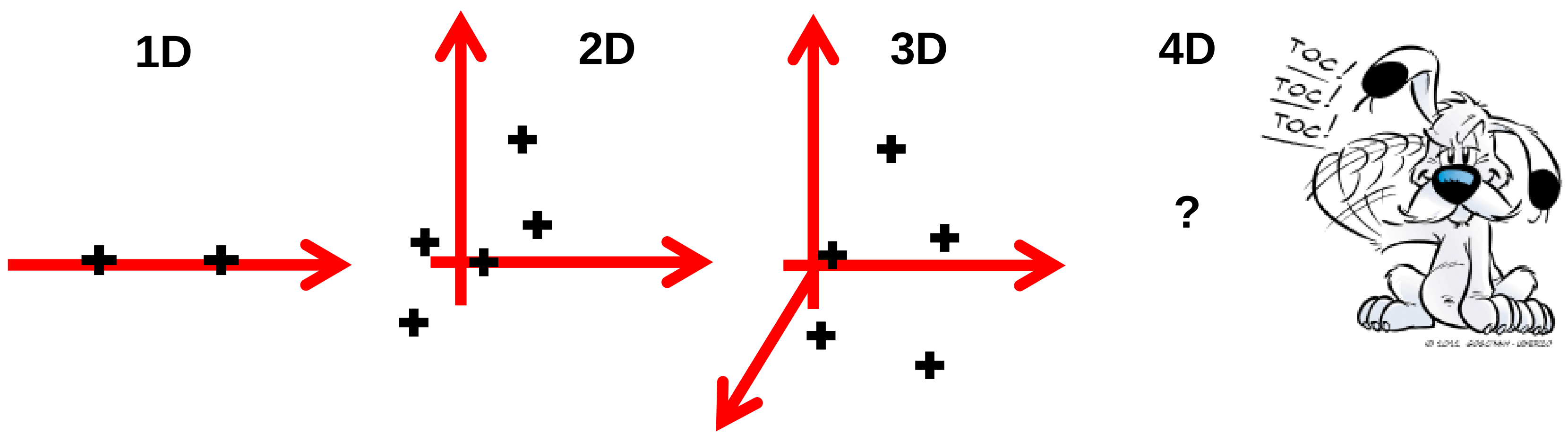
Source separation



PCA in Machine Learning

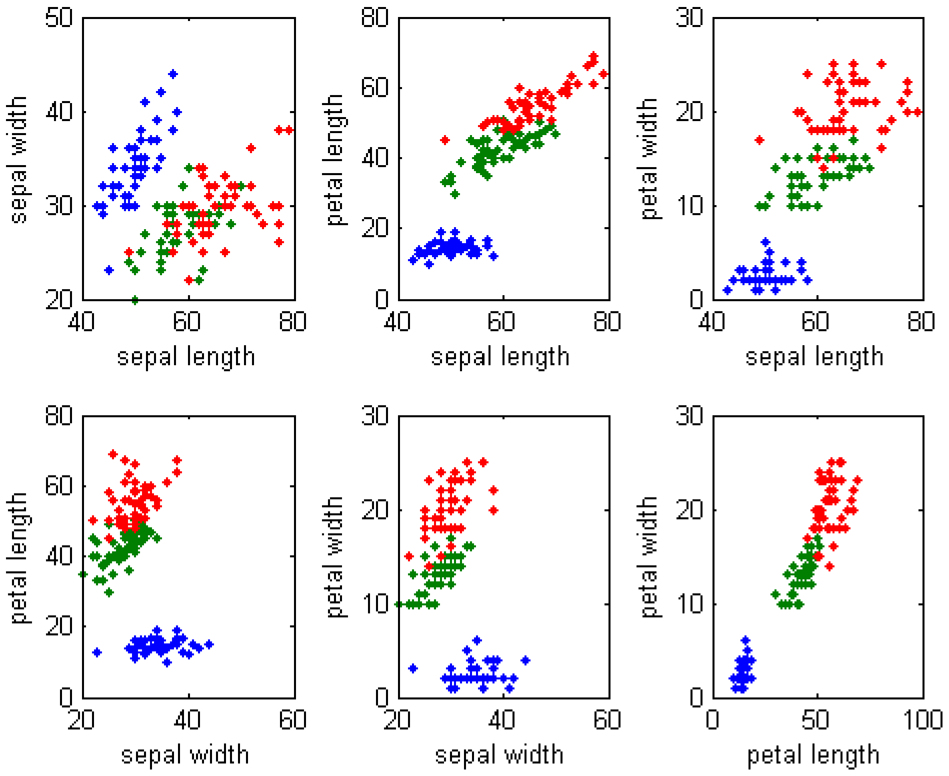
PCA usage in Machine Learning

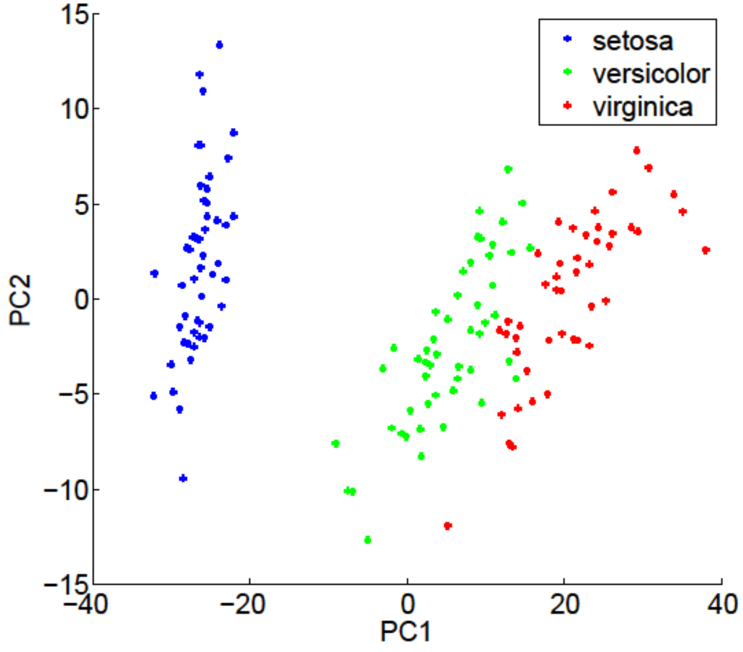
- We introduced PCA in the context of BSS which looks to separate the recorded data into their original sources.
- These techniques can be used for other purposes while keeping the exact same concept of "**finding a new basis** that *better* describes the underlying data":
  - **Visualization:** how to visualise data in $\mathbb{R}^n, n > 3$??
  - **Dimensionality reduction:** ML and dimensionality curse, remove redundancy.



PCA and ML

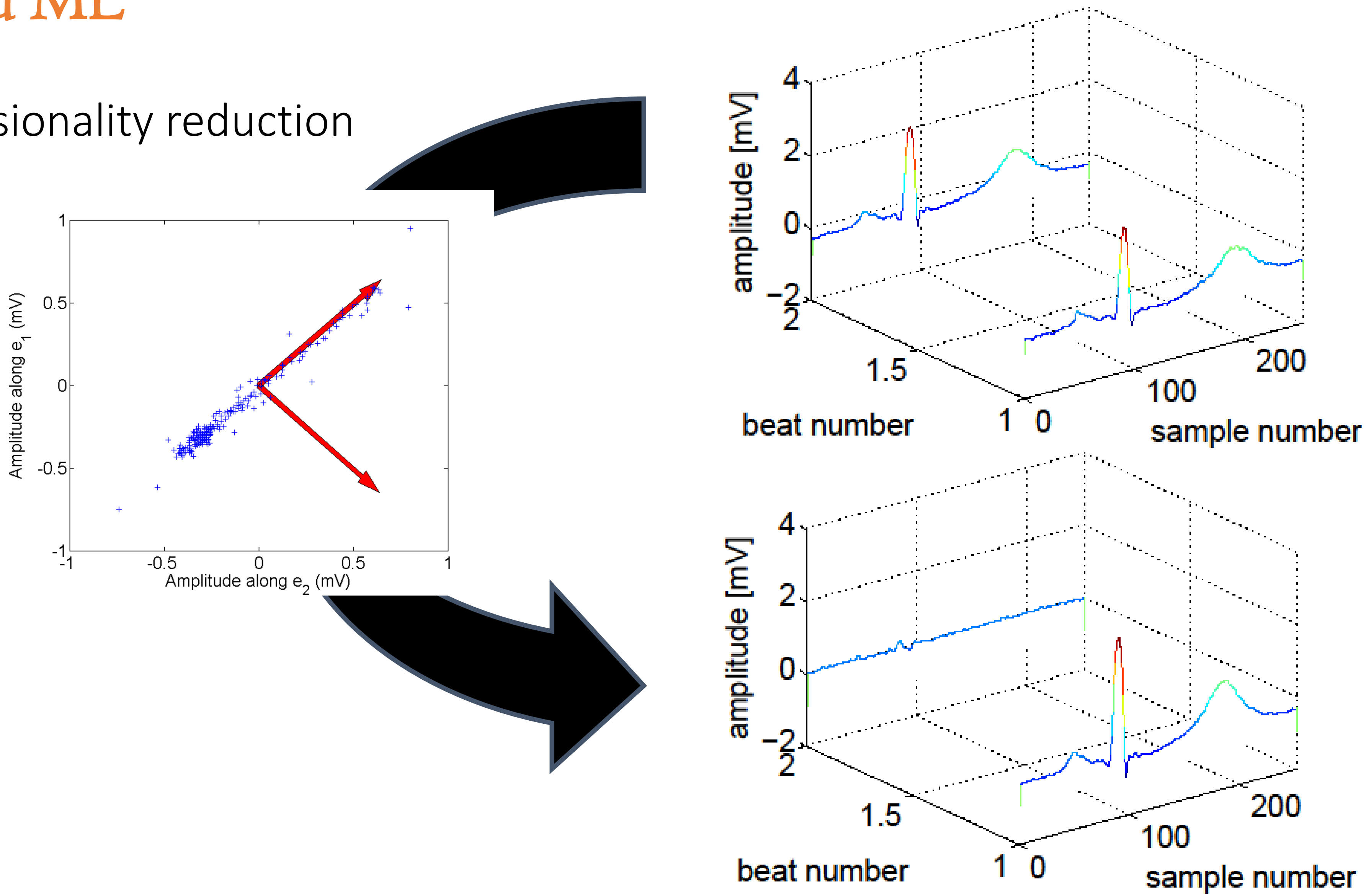
- **Visualization** of data in large dimensional space or **reduction of the number of features** to avoid **curse of dimensionality**.





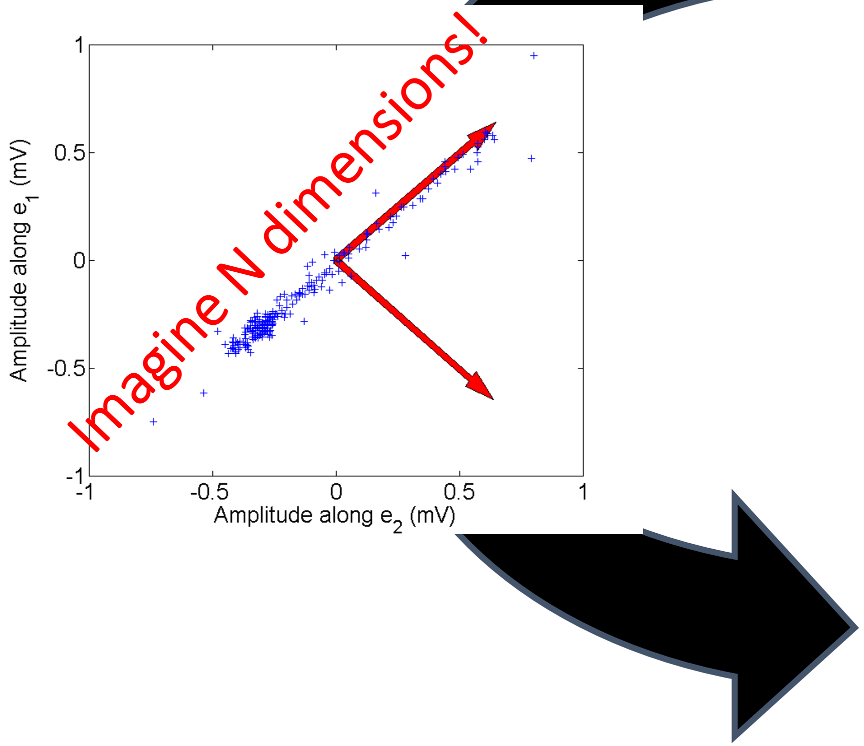
PCA and ML

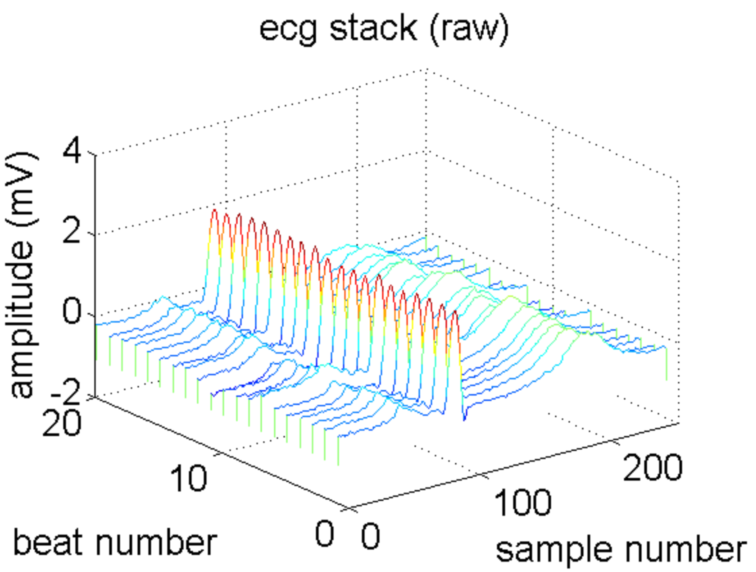
- Dimensionality reduction

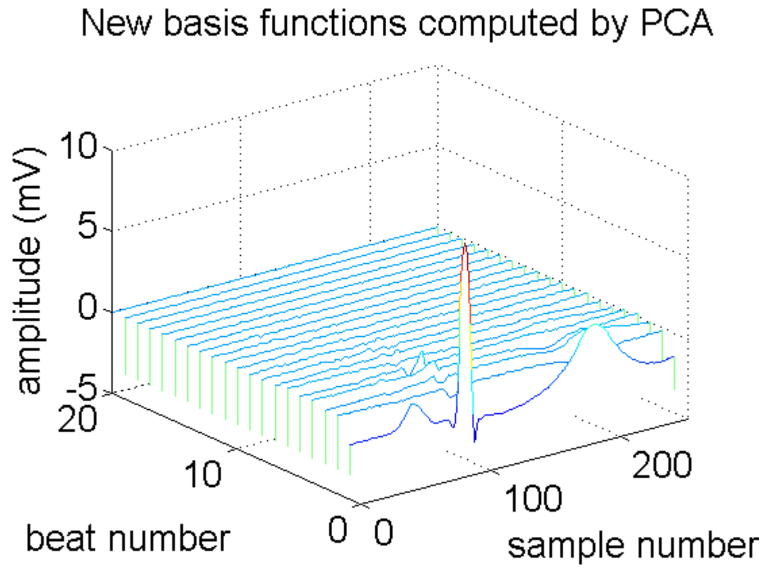


PCA and ML

- Dimensionality reduction



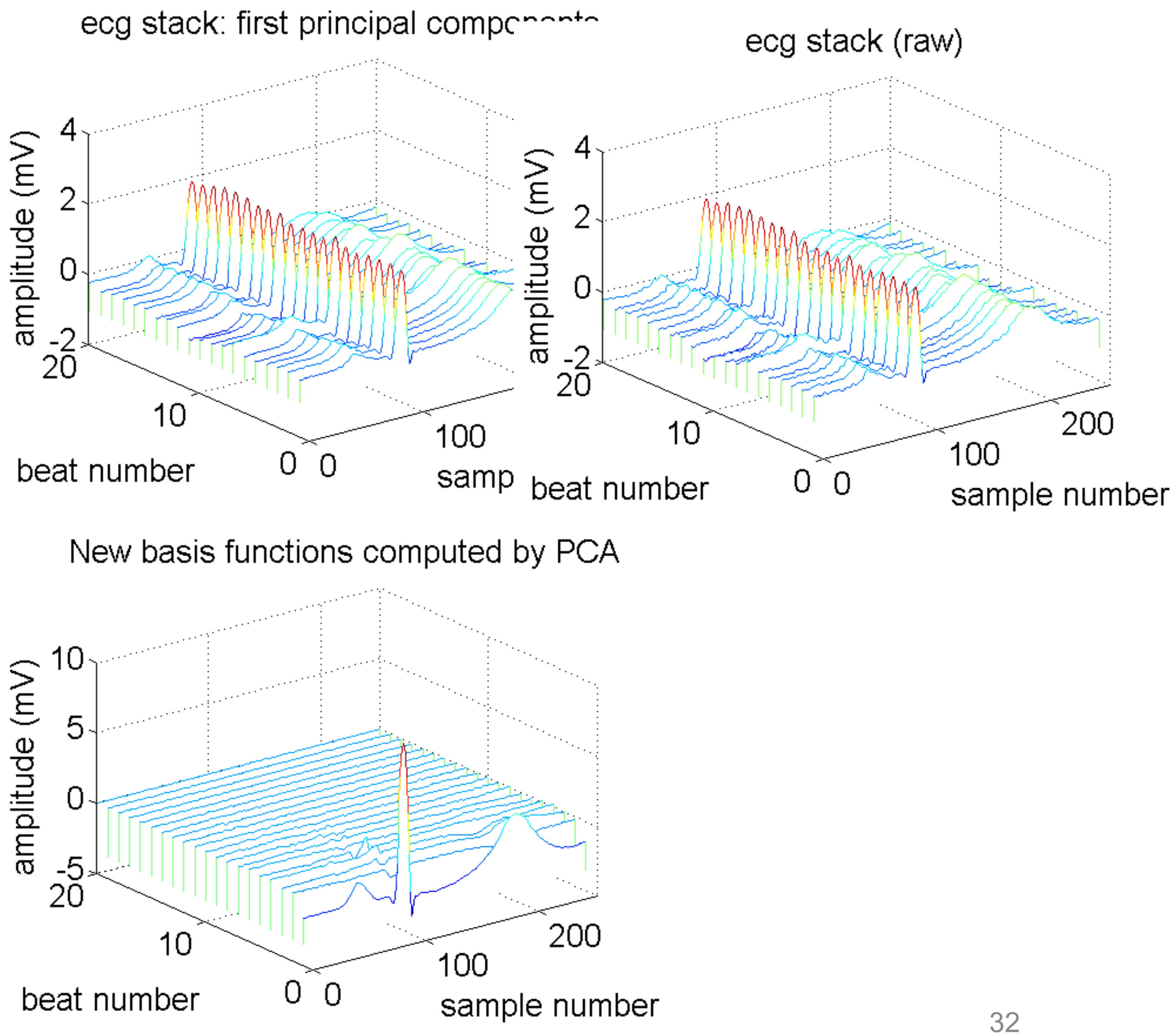




PCA and ML

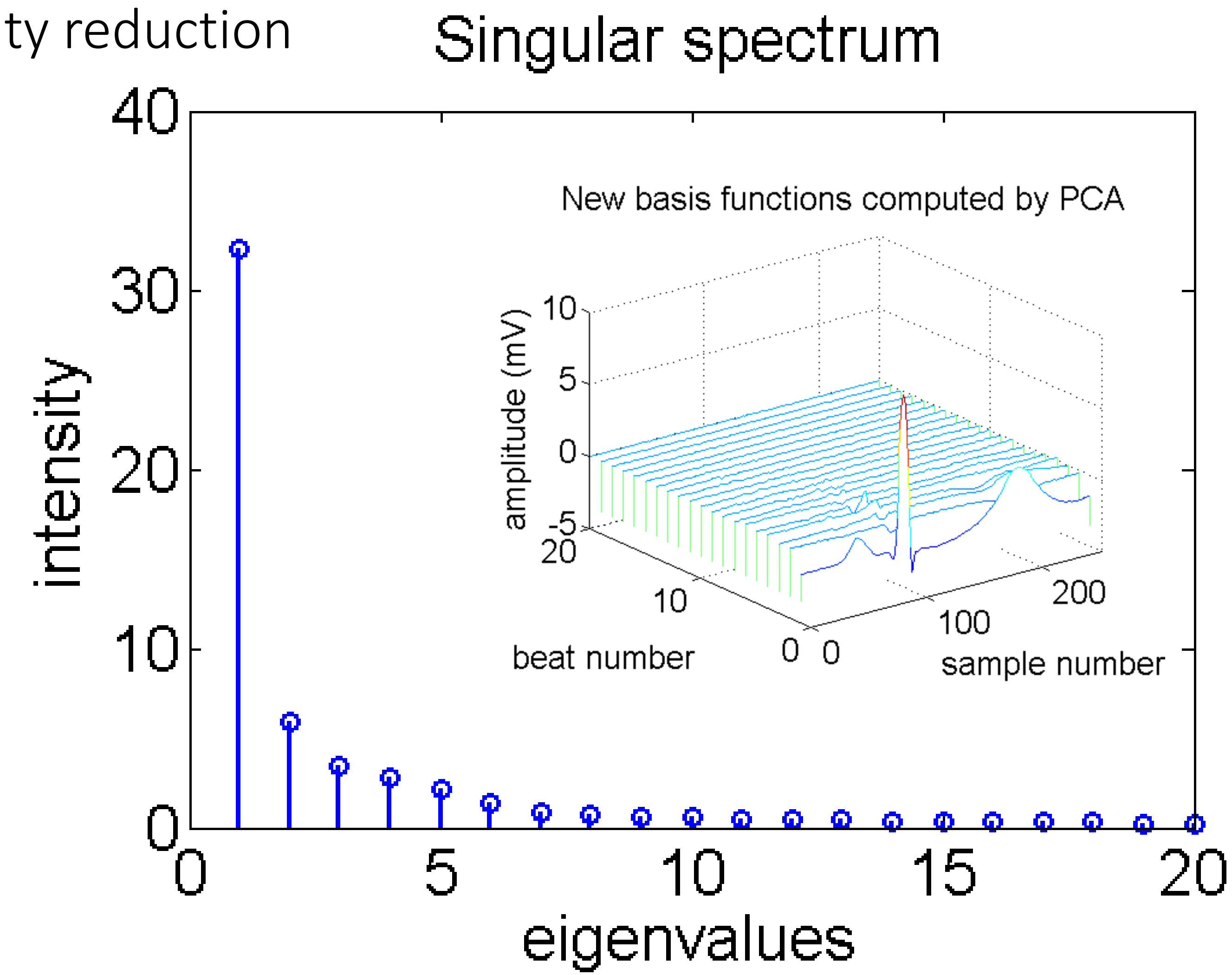
- Dimensionality reduction

Only keep first 4 dimensions
and project back



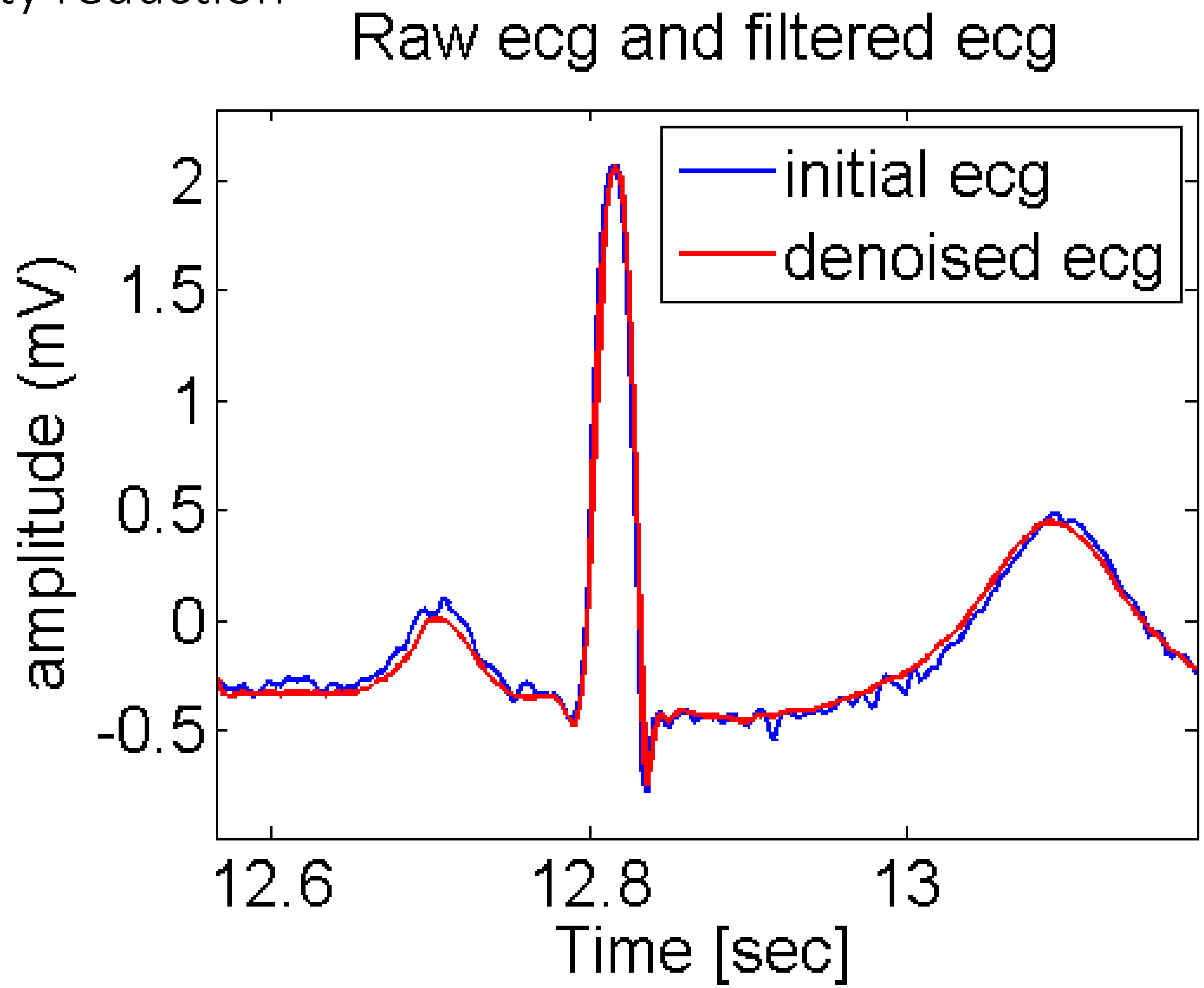
PCA and ML

- Dimensionality reduction



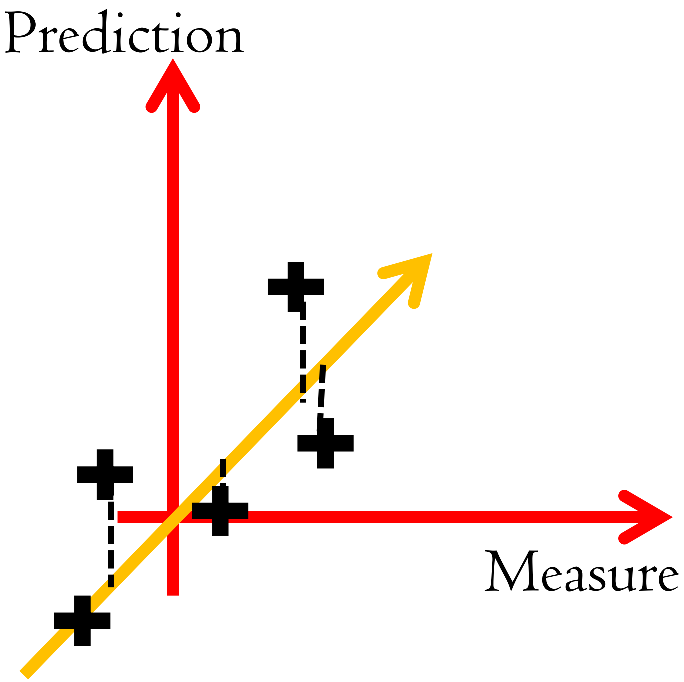
PCA and ML

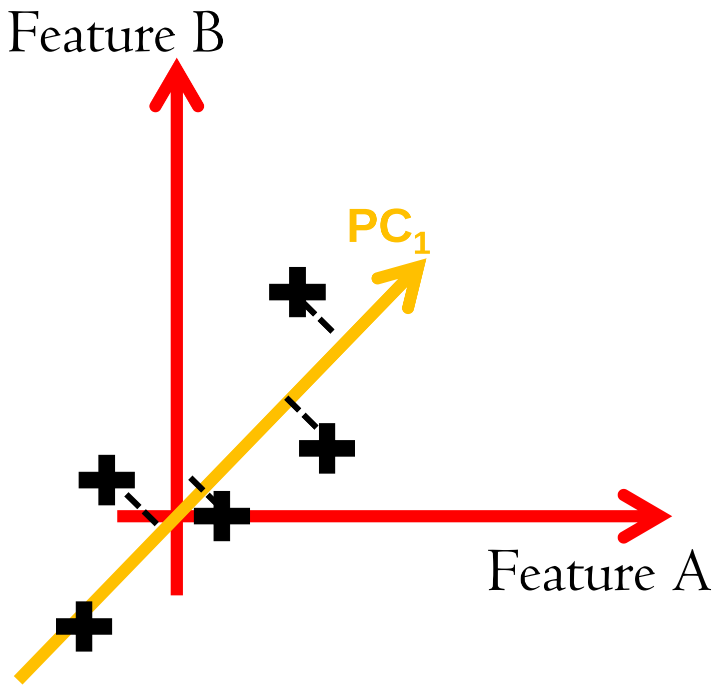
- Dimensionality reduction



Quiz: PCA and Linear Regression

- Is PCA equivalent to linear regression?

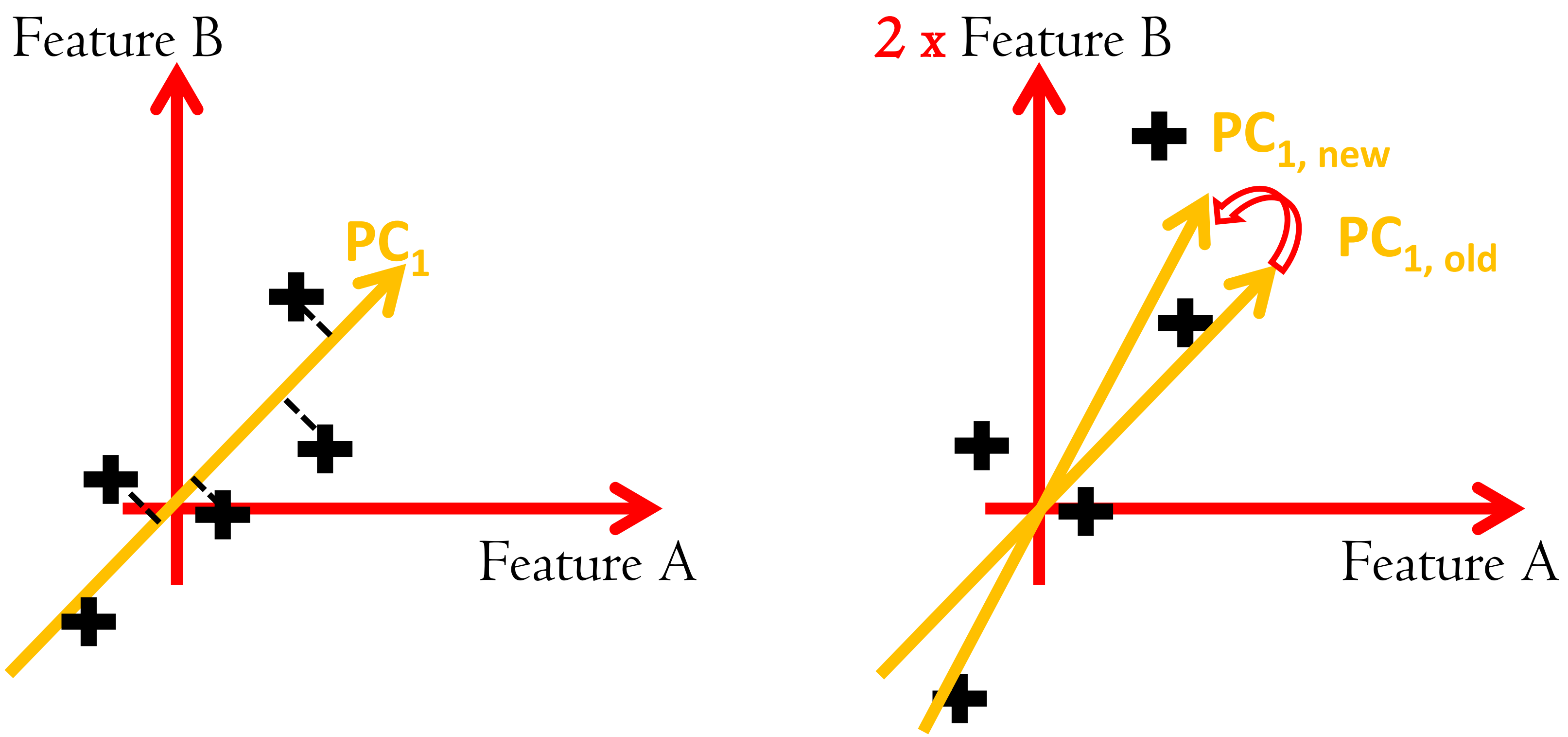




PCA is <u>NOT</u> linear regression

Quiz: PCA and Linear Regression

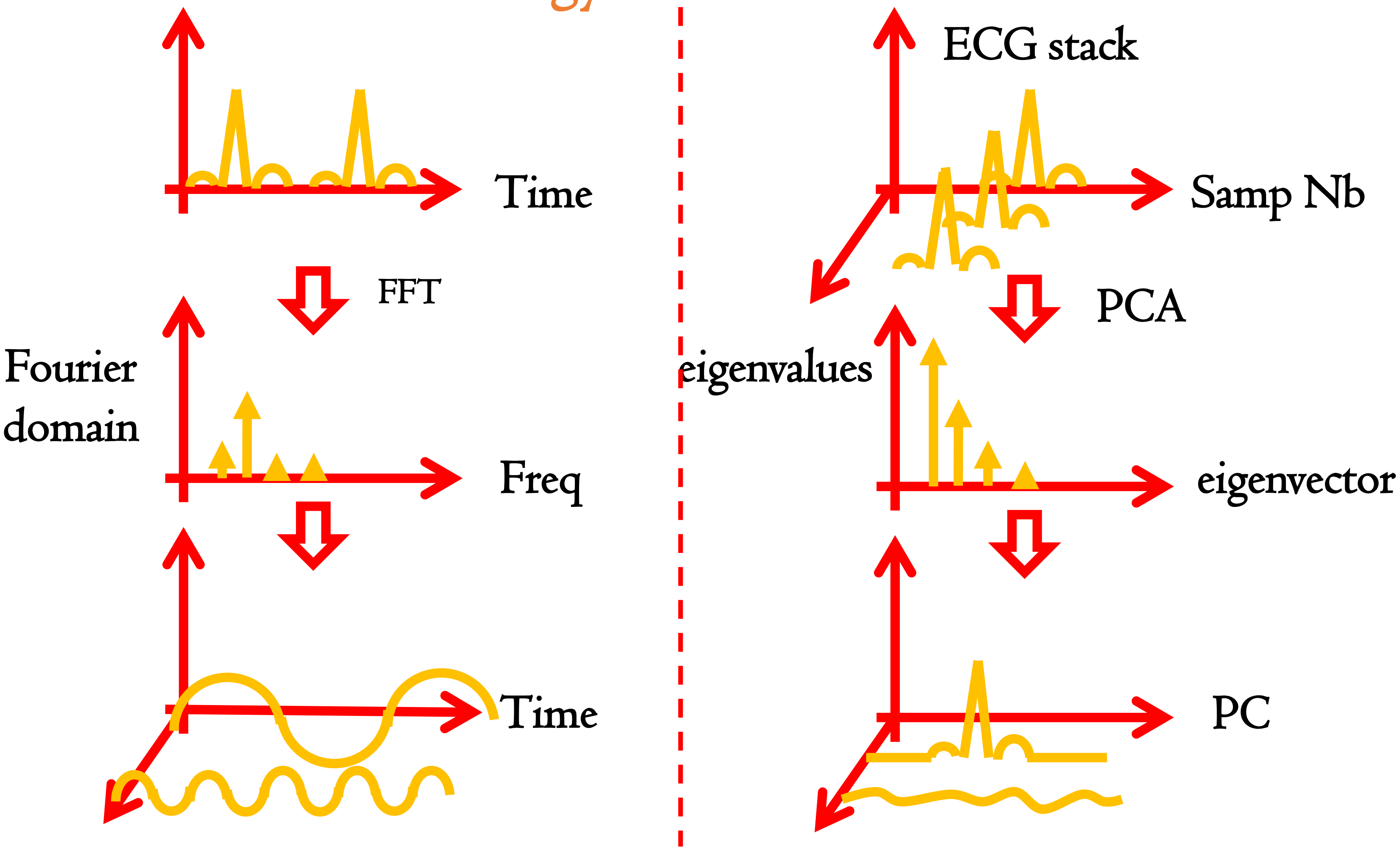
- Do we need to normalise the data?



PCA IS sensitive to scale

AIMLab.

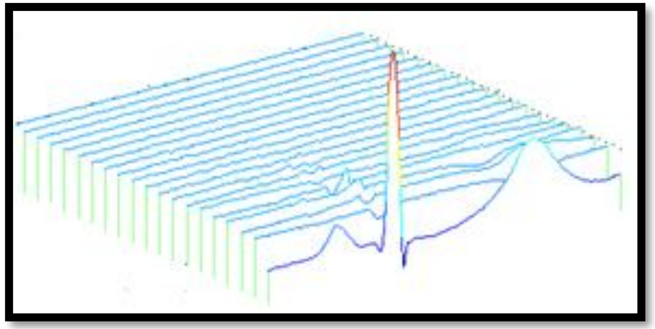
Quiz: PCA and Fourier Analogy



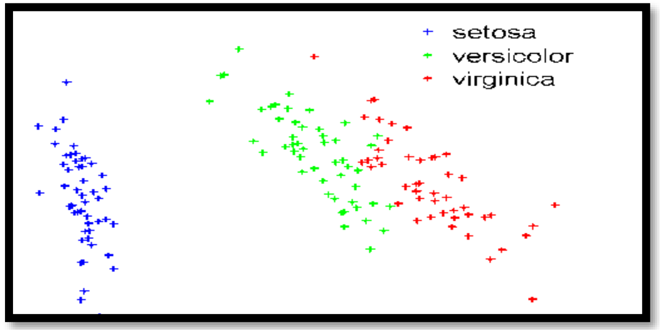
Summary

Summary

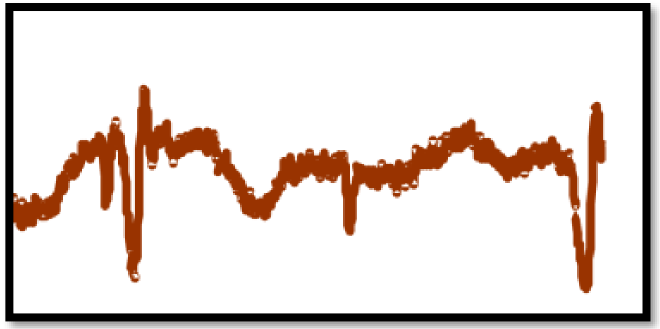
Dimensionality reduction



Visualisation



Source separation

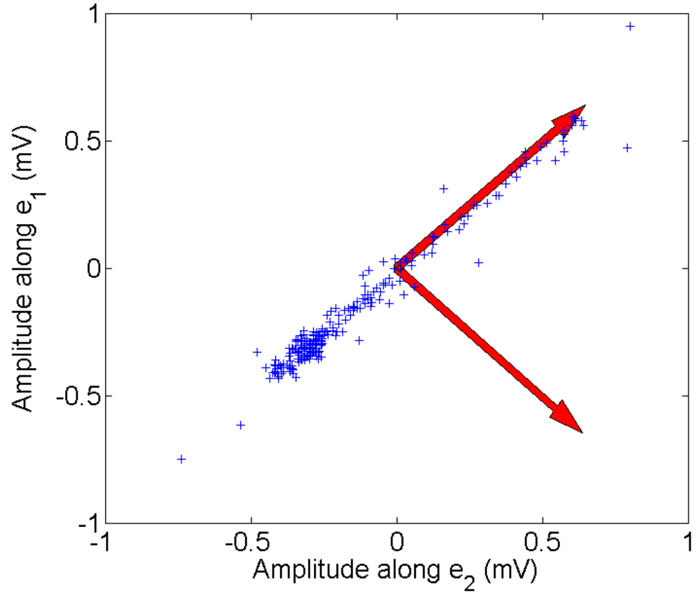


AIMLab.

Summary: Principal Component Analysis

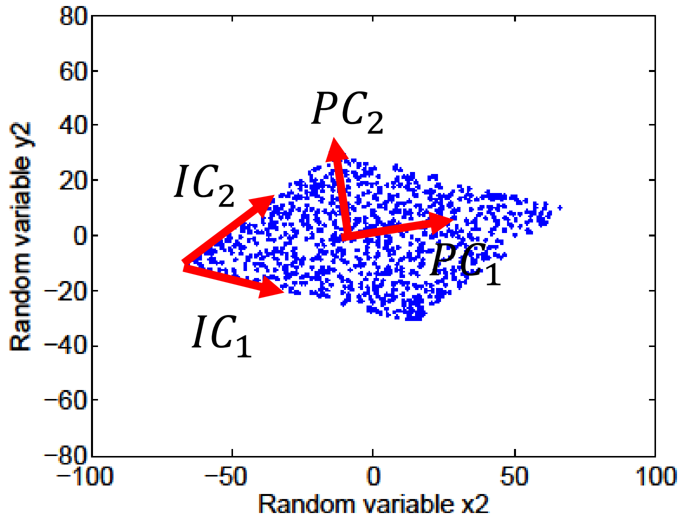
- Ideas we introduced here:
  - Expressing our dataset in a new basis may be a good idea!
  - PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

- Limitations with PCA:
  - Is maximal variance the right statistical criteria?
  - Limited to orthogonal basis. (Due to our criteria for independence which is second order.)



Summary: Limitations of PCA

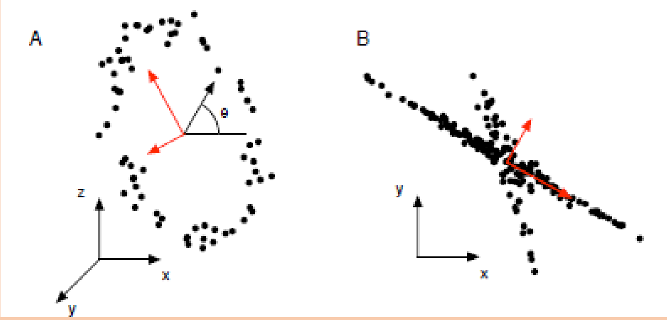
- As in PCA, we want to find a new vector basis on which to project our observations in order to obtain a set of maximally independent source signals.
- Instead of using variance as our independence measure (i.e. decorrelation) as in PCA, we will look for **statistical independence** with ICA.



Note, invention of ICA: Herault, Jeanny, and Christian Jutten. "Space or time adaptive signal processing by neural network models." Neural networks for computing. Vol. 151. No. 1. AIP Publishing, 1986.

Take Home

- BSS aims to recover the original **independent** sources from the observed **linear** mixtures.
- There are different ways of expressing “independence”.
- One way with PCA is to assume that **large variance** represent interesting structures.
- PCA is simple and a **non-parametric** method (unsupervised learning). It provides an analytical solution to the problem.
- PCA is constrained to orthogonal axes and defining large variance as independence is limiting.
- Way to deal with these limitations:
  - Kernel PCA.
  - ICA.



AIMLab.

References

[1] Gari D. Clifford course note: http://www.mit.edu/~gari/teaching/6.555/SLIDES/BSShandouts.pdf
[2] Joachim Behar. Course note, ML in Healthcare course.
[3] Independent Component Analysis: Algorithms and Applications. Aapo Hyvärinen and Erkki Oja. URL: http://mlsp.cs.cmu.edu/courses/fall2012/lectures/ICA_Hyvarinen.pdf
[4] ICA for dummies. Online tutorial by Arnaud Delorme. URL: http://sccn.ucsd.edu/~arno/indexica.html
[5] Clifford, Gari D., and Francisco Azuaje. Advanced methods and tools for ECG data analysis. London: Artech house, 2006.
[6] Aapo Hyvarinen. tutorial on whitening. URL: http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/node26.html
[7] Cardoso, J-F. "Blind signal separation: statistical principles." Proceedings of the IEEE 86.10 (1998): 2009-2025.
[8] Course notes: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf
[9] Course notes: http://cis.legacy.ics.tkk.fi/aapo/papers/NCS99web/node33.html
[10] Arnaud Delorme research page: http://arnauddelorme.com/ica_for_dummies/