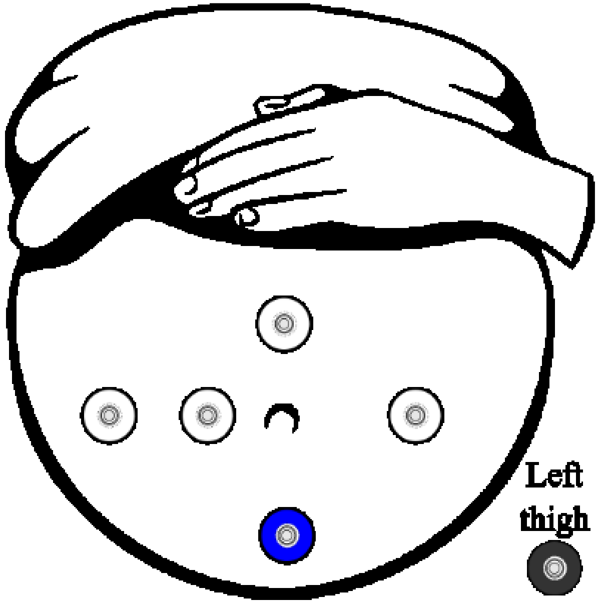
Machine Learning in Healthcare

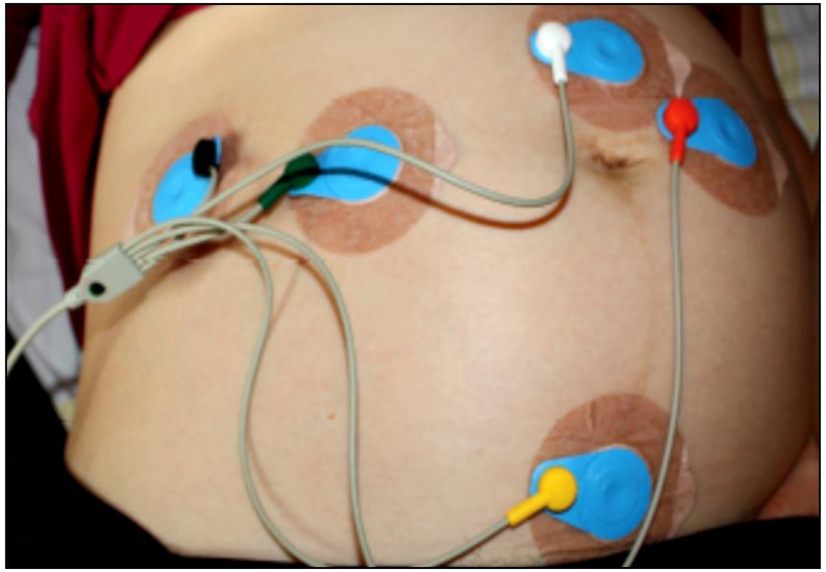
#L14-Independant component analysis

Technion-IIT, Haifa, Israel

Asst. Prof. Joachim Behar
Biomedical Engineering Faculty, Technion-IIT
Artificial intelligence in medicine laboratory (AIMLab.)
https://aim-lab.github.io/
Twitter: @lab_aim

NI-FECG



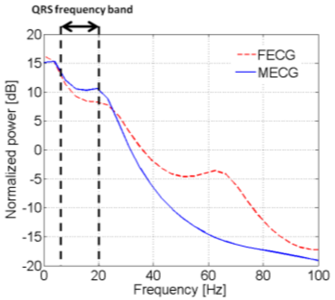


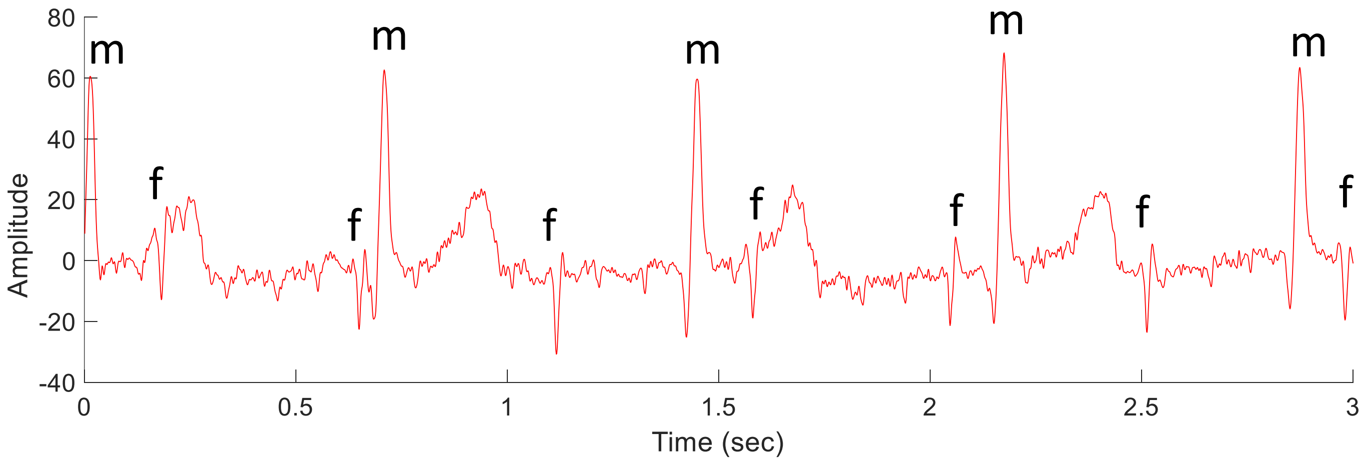
NI-FECG: opportunity

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

NI-FECG: Challenges

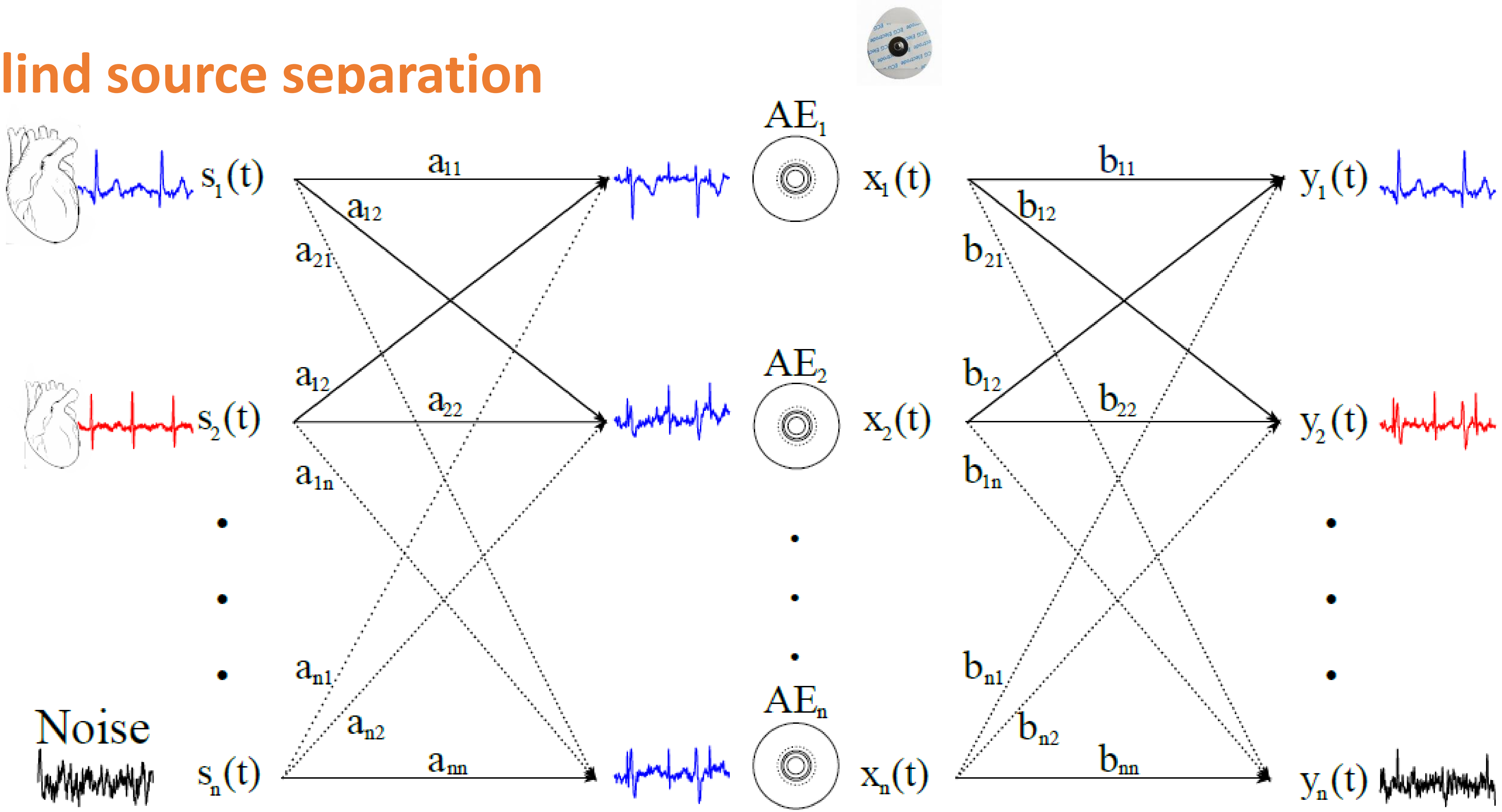
- Overlap in time and frequency,
- Non stationarities,
- Vernix caseosa.





Reminder

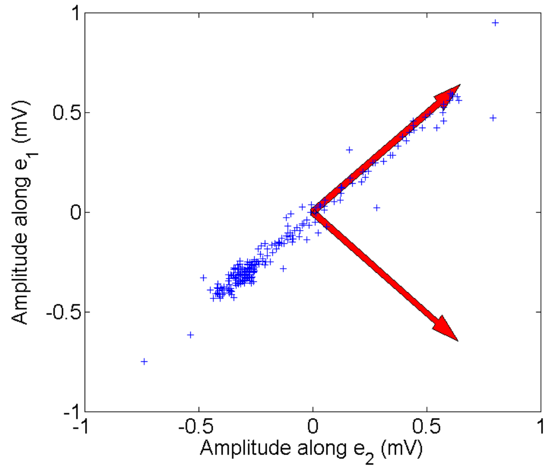
Blind source separation



AIMLab.

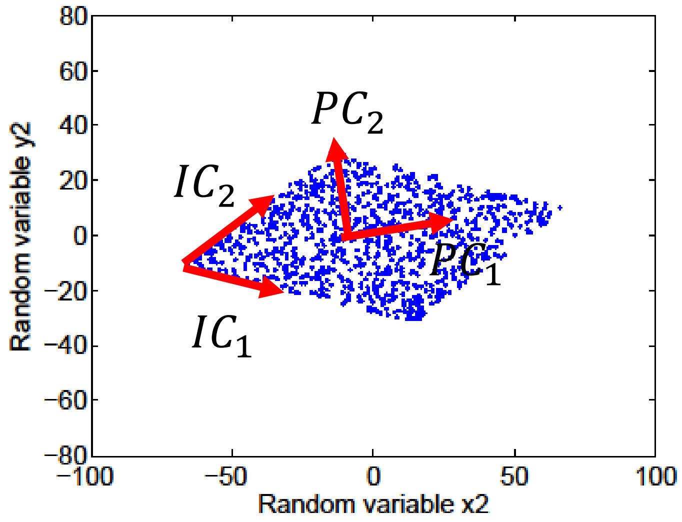
Principal component analysis

- 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.)



Independent Component Analysis

- 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 **<u>statistical independence</u>** with ICA.



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.

Independent component analysis

Independent component analysis

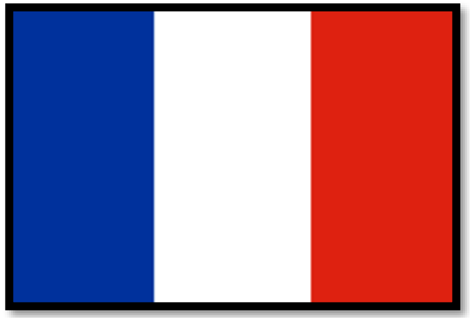

Codename

Herault and Jutten, 1986

Special power

Source separation

Place of origin



ICA

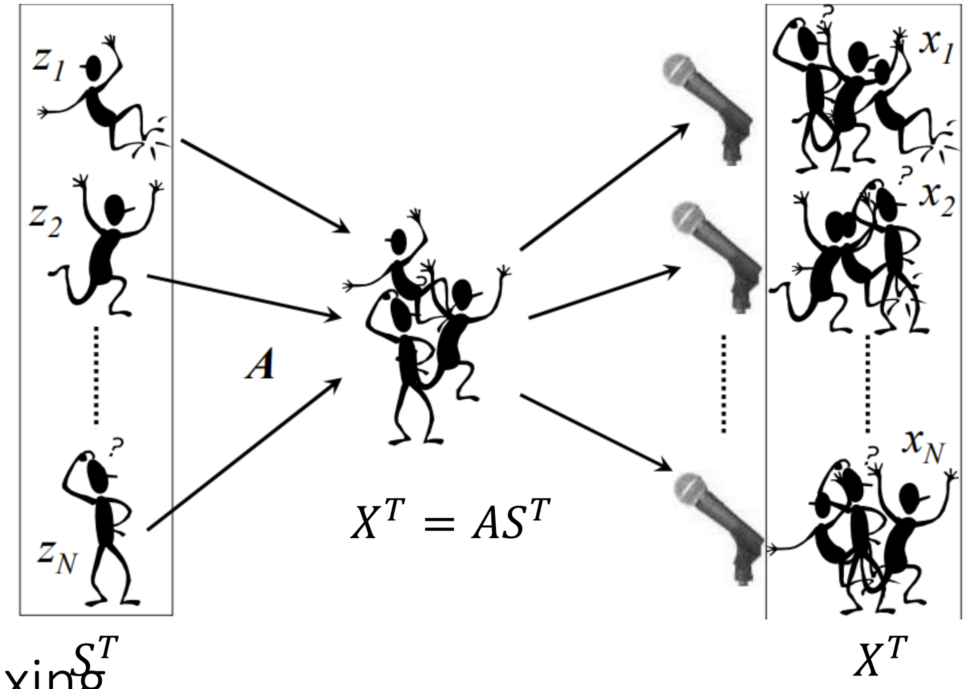
Herault, Jeanny, and Christian Jutten. "Space or time adaptive signal processing by neural network models." Neural networks for computing. Vol. 151. No. 1. AIP Publishing, 1986.

Independent component analysis

- Independent Component Analysis (ICA) consists of recovering unobserved signals or sources from several observed mixture by exploiting the assumption of **mutual independence** between the signals [Card1998].
- Two microphones recording two individuals:
  - $\begin{cases} x_1(t) = a_{11}s_1 + a_{12}s_2 \\ x_2(t) = a_{21}s_1 + a_{22}s_2 \end{cases}$
  - More generally we write: $x = As$ and $s = Wx$
  - A is commonly called the **mixing matrix**.
  - We assume that $x_1(t)$ and $x_2(t)$ are **linear** and **instantaneous mixtures**.

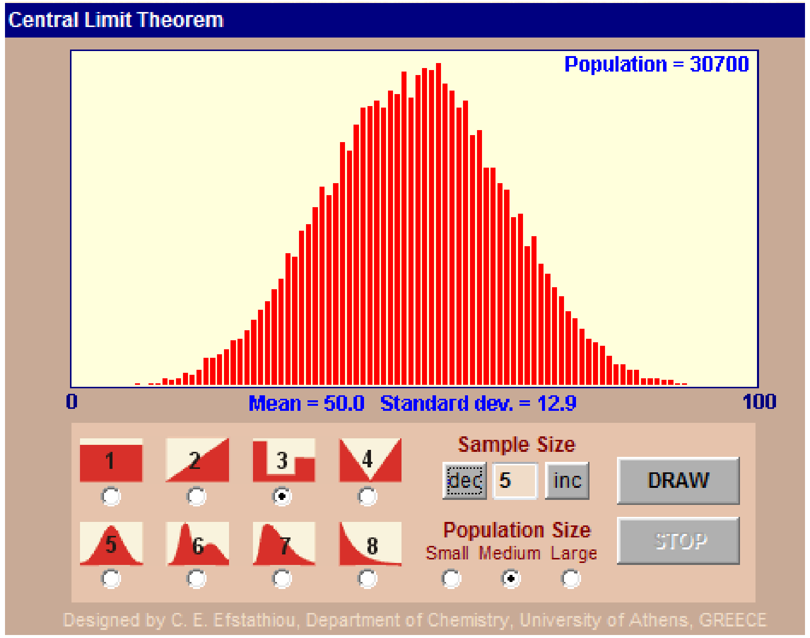
The Cocktail Party Problem

- 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 $W = \hat{A}^{-1}$.
- Iterative process with some **cost function** which measures the statistical independence of the estimated sources at each iteration.



Non-Gaussianity as statistical independence

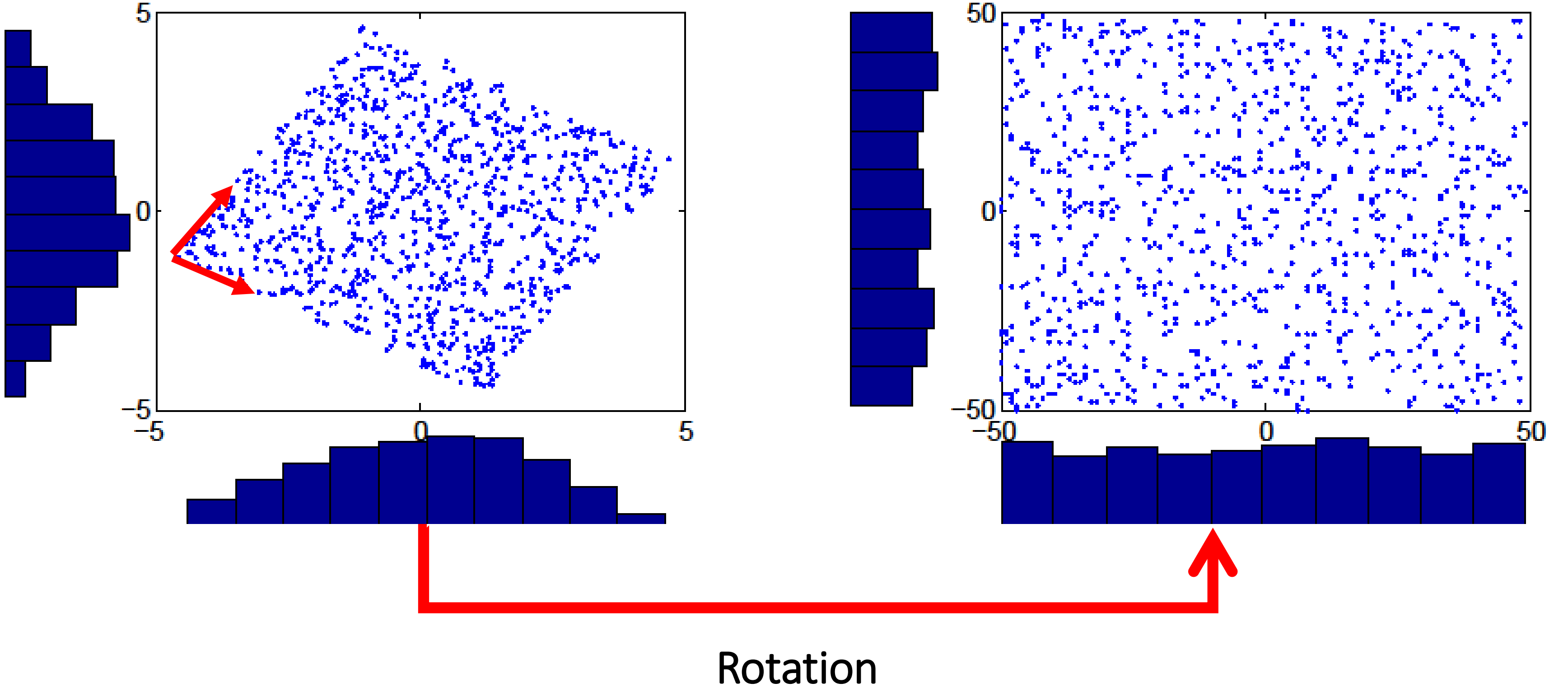
- Reminder, **Central limit theorem**: when independent random variables are added, their properly normalized sum tends toward a normal distribution.
- This is true, whatever the type of distributions of the individual random variables.
- Thus to demix our signal we will look for maximize non-Gaussianity. Non-Gaussianity is our marker of independence.



Example: http://www.chem.uoa.gr/applets/appletcentrallimit/appl_centrallimit2.html

Non-Gaussianity as statistical independence

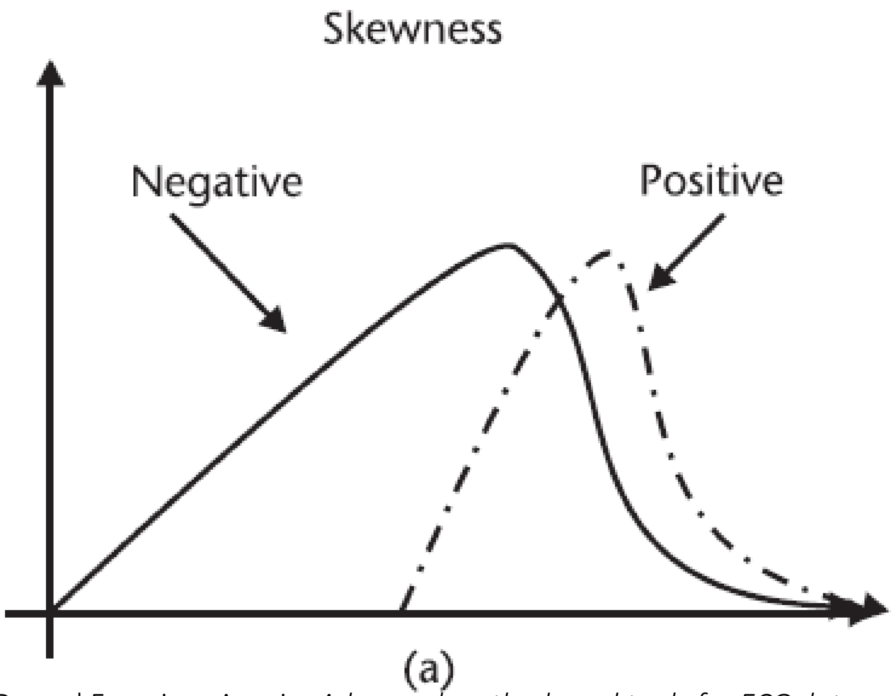
- Central limit theorem: any linear mixture of 2 i.i.d. random variables is more Gaussian than the original variables.



Higher order moments

- A moment is a specific quantitative measure of the shape of a probability distribution.
- 2nd order is variance (we used it in PCA), 3rd moment skewness and 4th moment is kurtosis.

$$s(x) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{x^{(i)}-\mu}{\hat{\sigma}}\right]^3 \qquad k(x) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{x^{(i)}-\mu}{\hat{\sigma}}\right]^4$$



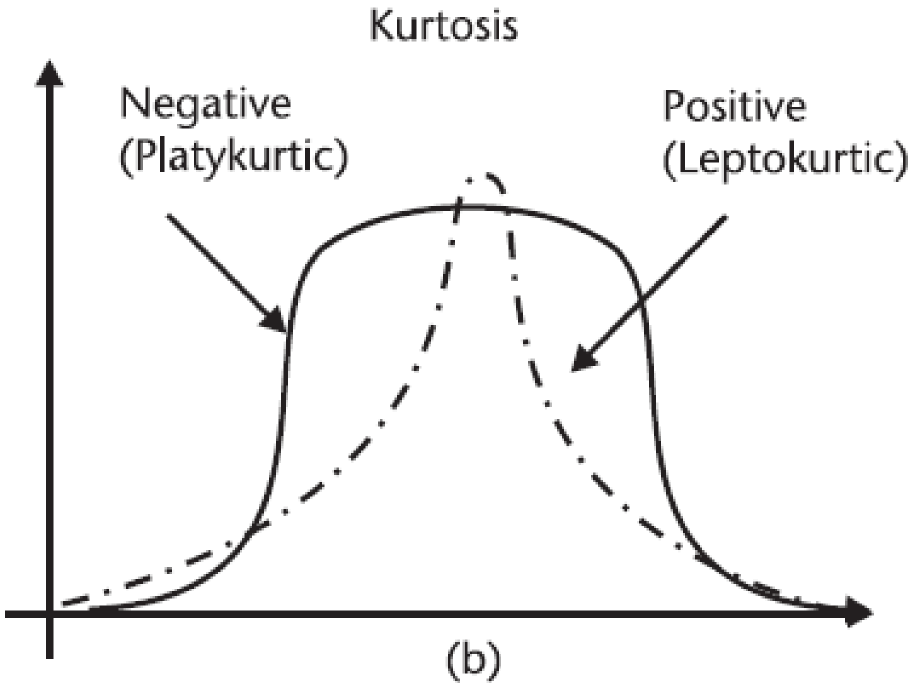


Image: Clifford, Gari D., and Francisco Azuaje. *Advanced methods and tools for ECG data analysis*. London: Artech house, 2006.

Independent component analysis

- So we can use kurtosis for example as the cost function and proceed:
  - Initialize W.
  - Iterate to update W with gradient descent to maximize kurtosis.
- Non-Gaussianity is one approximation, but sensitive to small changes in the distribution tail.
- Other measures of statistical independence may be used.

(Recall: Higher order moments $E(X^n) = \int_{-\infty}^{+\infty} x^n f(x)dx$. Mean, variance, skewness, kurtosis, hyperskewness, hyperflatness!)

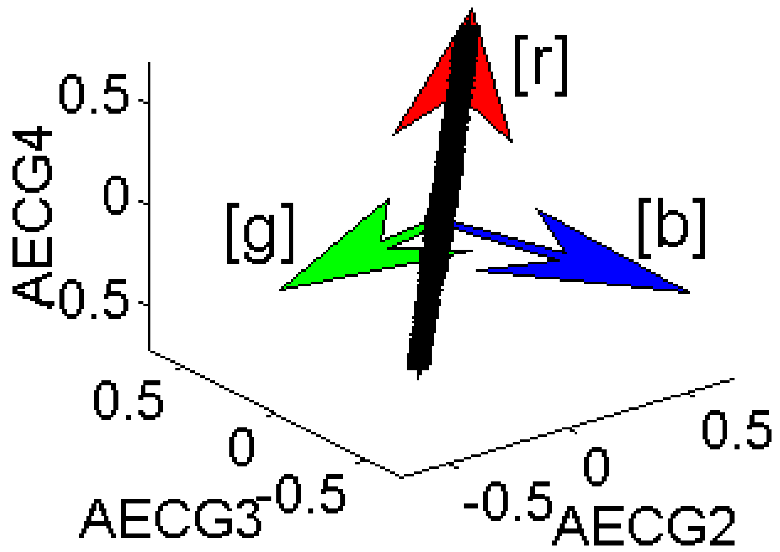
Independent component analysis

- Looking for higher order statistics:
  - Kurtosis: $kurt(y) = E\{y^4\} - 3(E\{y^2\})^2$
  - Negentropy approximated: $J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2.$
  - Mutual information
  - Cummulants (JADE)
  - …
  - These are called **contrast function**. Their optimization allows to estimate the independent components.

(Recall: Higher order moments $E(X^n) = \int_{-\infty}^{+\infty} x^n f(x) dx$. Mean, variance, skewness, kurtosis, hyperskewness, hyperflatness!)

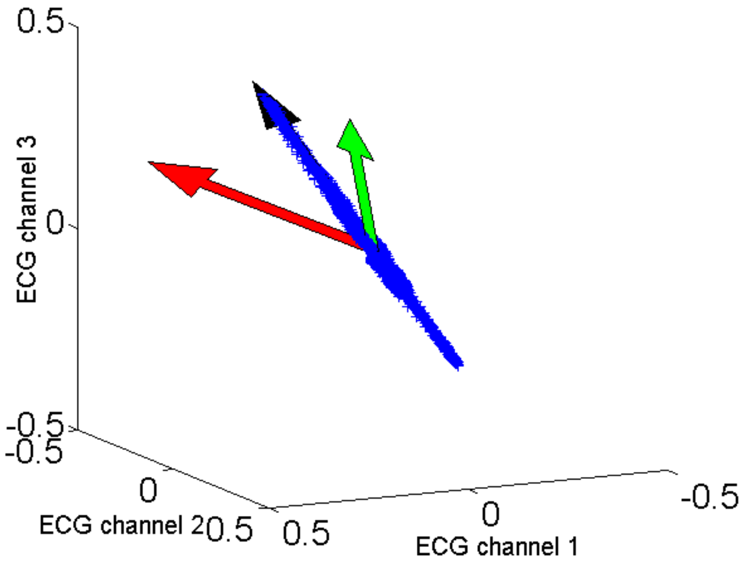
Independent component analysis

<u>PCA</u>

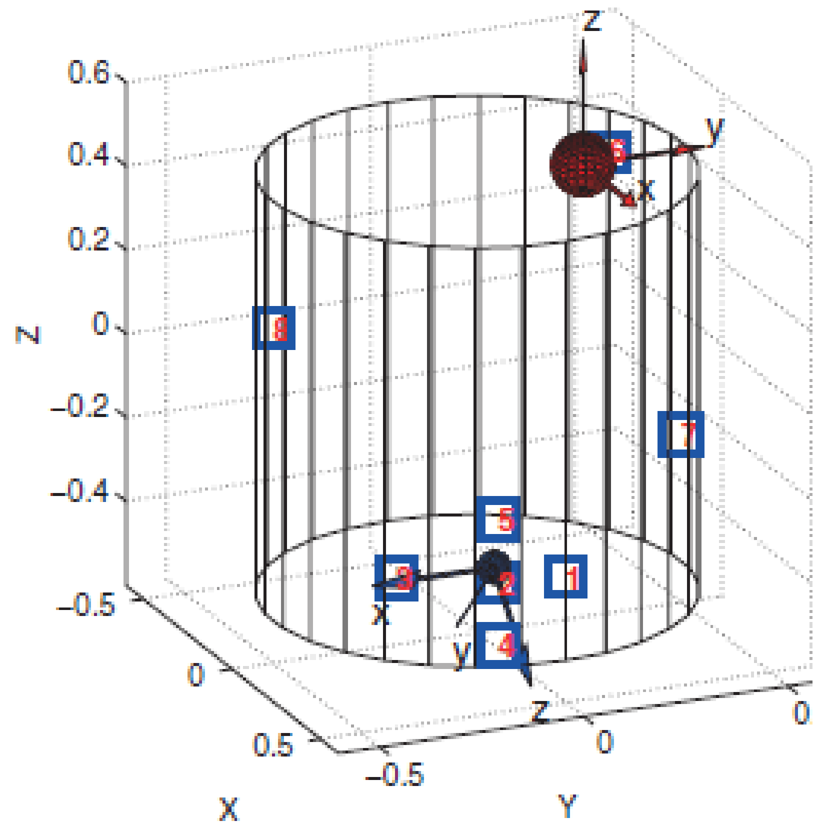


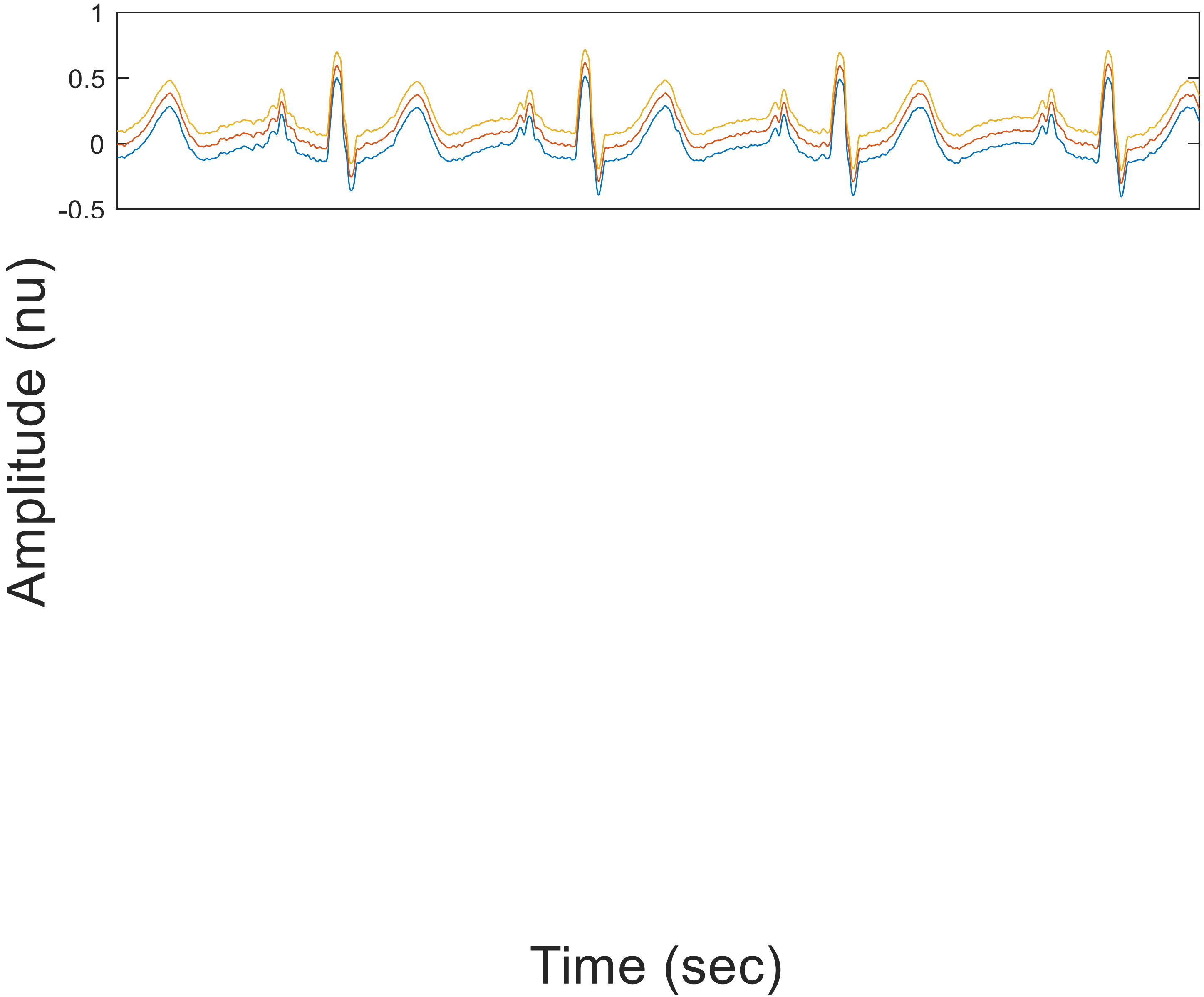
- Maximal covariance,
- Closed form solution,
- Constraint to orthogonal axis.

<u>ICA</u>



- Statistical independence,
- No closed form solution,
- Not constraint to orthogonal axis.


0.6
0.4
0.2
0
−0.2
−0.4
−0.5
0
0.5
Z
X
Y
−0.5
0
0
1
2
3
4
5
6
7
8
x
y
z

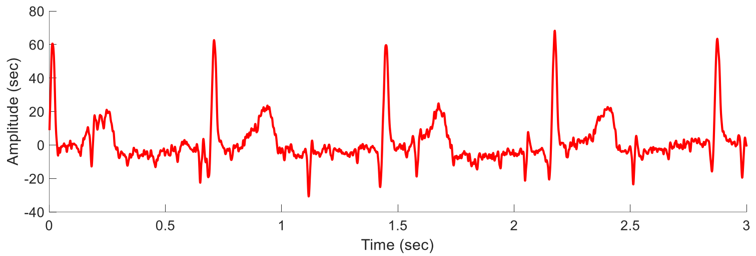

1
0.5
0
-0.5
Amplitude (nu)
Time (sec)

Independent component analysis

- In summary:
  - ICA aims to find a linear representation of nongaussian data so that the components are statistically independent.
  - ICA exploits the spatial diversity. Time structure are ignored. The mixture is assumed to be instantaneous.
  - ICA assumes, independence of the sources (and that the sources are not Gaussians). It does not assume orthogonality of the axis.
- Limitation:
  - Only works for linear mixture.
- Example of good ICA implementations: JADE, FastICA.

AIMLab.

Algorithm evaluation

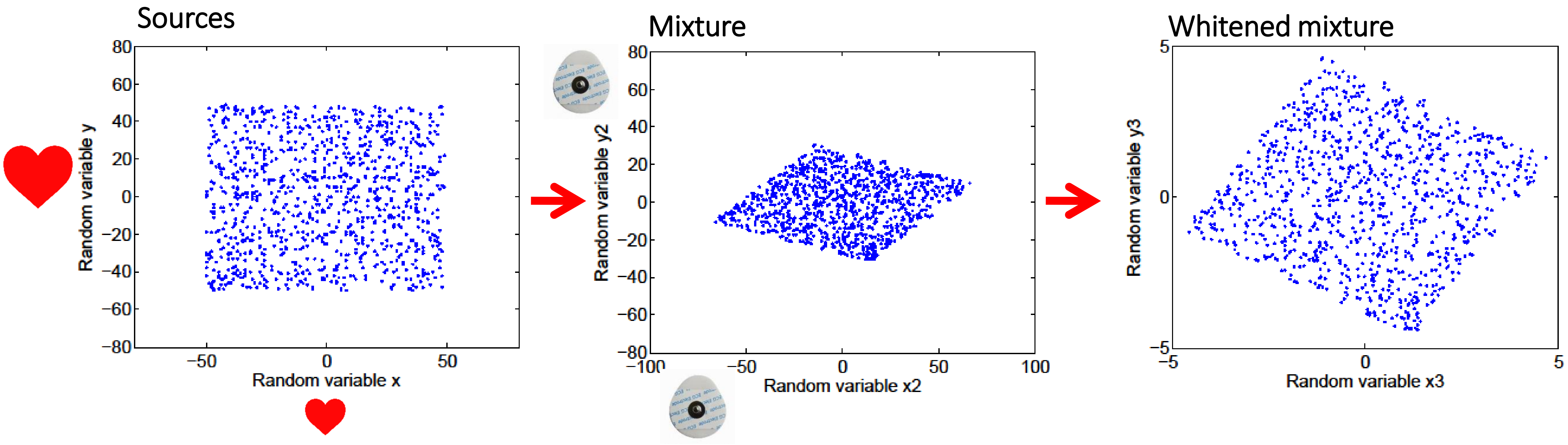


CL	Method	HRE	RRE	Se	PPV	F_1	$F_1\dagger$
		NU	NU	%	%	%	%
I	TS	655.5	27.9	81.8	81.7	81.6	81.2
I	TS_c	514.8	29.1	81.6	81.7	81.5	81.4
I	TS_m	551.9	28.1	82.2	82.2	82.1	81.1
I	TS_{lp}	902.0	46.2	82.1	81.9	81.8	78.5
I	TS_{pca}	594.4	21.6	88.1	84.5	86.1	83.6
I	TS_{EKF}	733.8	25.0	83.0	81.1	81.9	78.4
II	ICA	2852.1	39.3	69.1	60.0	63.7	61.7
II	PCA	3892.1	45.3	57.4	47.9	51.6	52.6
III	TS-ICA	272.7	17.1	93.0	91.1	92.0	91.3
III	TS_c-ICA	202.6	17.2	93.2	92.0	92.6	92.1
III	TS_m-ICA	251.9	18.4	91.7	90.8	91.2	92.3
III	TS_{lp}-ICA	399.1	37.9	88.4	88.7	88.4	85.2
III	TS_{pca}-ICA	153.2	16.9	93.8	92.2	93.0	92.4
IV	ICA-TS_{pca}	396.9	27.1	90.1	88.8	89.2	89.4
IV	ICA-TS_{pca}-ICA	299.4	22.7	92.6	92.2	92.4	91.1
	CONST-HR (143 bpm)	172.2	8.9	23.2	23.1	23.0	NA
	FUSE	132.9	12.7	95.6	94.3	95.0	94.2
	FUSE-SMOOTH	19.1	6.3	95.9	96.0	96.0	95.2
	FUSE-CHALL	5.4	2.3	NA	NA	NA	NA

Behar et al. Physiol. Meas. 2014a

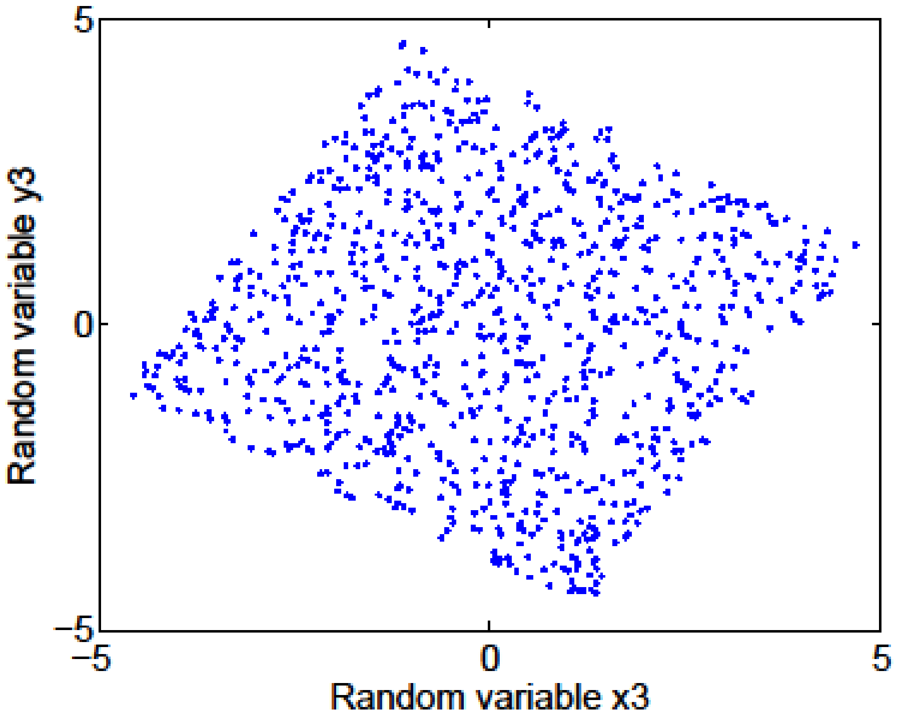
Extra insights on ICA

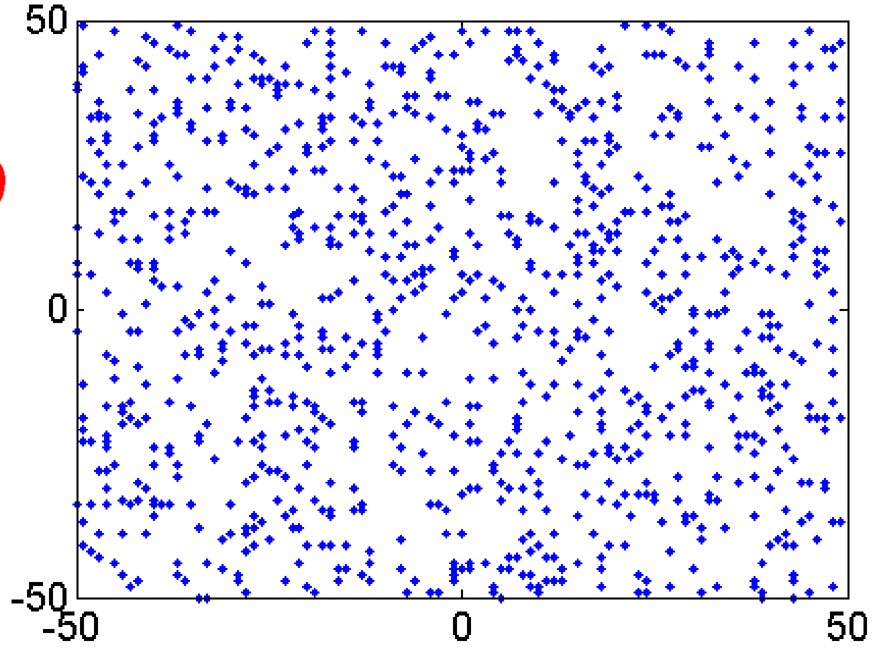
ICA: whitening the data



- **Whitening** removes any correlation in the data.
- The geometric interpretation is that whitening restores the initial shape of the data then ICA `only’ has to rotate.

ICA: whitening the data

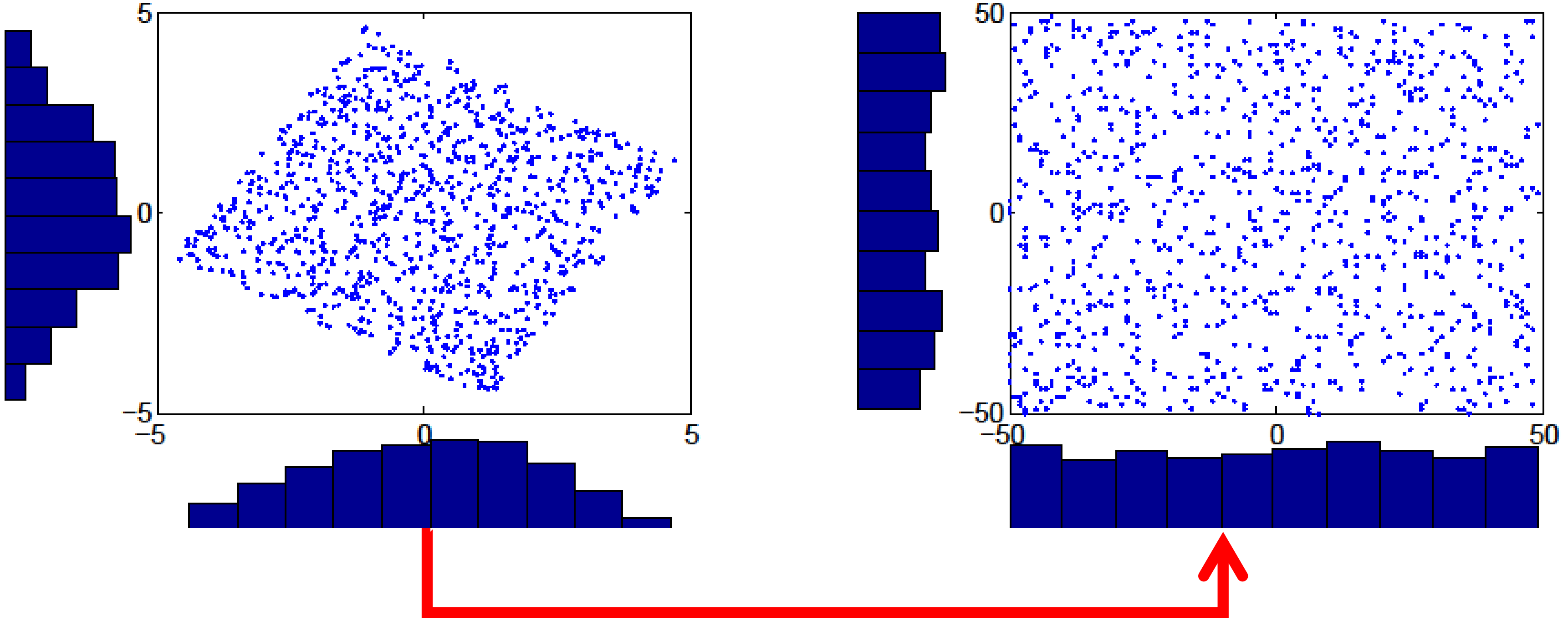




Rotation

ICA: rotation

- Central limit theorem: any linear mixture of 2 i.i.d. random variables is more Gaussian than the original variables.



Rotation

ICA: whitening the data

- Whitening procedure: we want to transform the matrix X linearly so that we obtain a new observation matrix X_w which is white i.e. its components are **uncorrelated** and with **unit variance** i.e. $E\{{X_w}^T X_w\} = I$ i.e. we want the covariance matrix to be the unity matrix.
- How can we produce this transform?
- Covariance matrix can be diagonalized (recall: we saw that in PCA)
  - $C = PDP^T$ where P is orthogonal.

ICA: whitening the data

- So an idea is that we can diagonalize first but then we need to normalize somehow to make the covariance matrix of unit variance:
  - $X_w = D^{-1/2}P^T X$
  - We can show that whitening is like transforming by $D^{1/2}P^T$
    - $X_w X_w^T = D^{-1/2}P^T X X^T P D^{-1/2} = D^{-1/2}P^T C P D^{-1/2} = I$

ICA: whitening the data reduces the complexity of the problem

- So we have:
  - $X_w = D^{-1/2}P^T X = D^{-1/2}P^T AS = A_{new}S$
  - A_{new} is orthogonal:
    - $X_w X_w^T = I = A_{new}S(A_{new}S\ )^T = A_{new}SS^T A_{new}^T$
    - We assume unit variance of the sources we look for,
    - $A_{new}SS^T A_{new}^T = A_{new}A_{new}^T = I$
  - A is an $N \cdot N$ matrix with N^2 degree of freedom.
  - A_{new} is an $N \cdot N$ matrix with $N \cdot (N-1)/2$ degree of freedom.
  - So by making **the whitening transformation we reduced by about half the complexity** of the problem. This is the purpose of whitening.

ICA algorithm

- Typical steps:
  - **Centering:** subtract the mean of the signal.
  - **Whitening:** uncorrelated the data. This means to treat all dimensions equally and simplify the ICA problem.
  - Dimensionality reduction (optional): remove PCA components with the least variance.
  - **Iterative algorithm:** find the independent components.

Independent Component Analysis

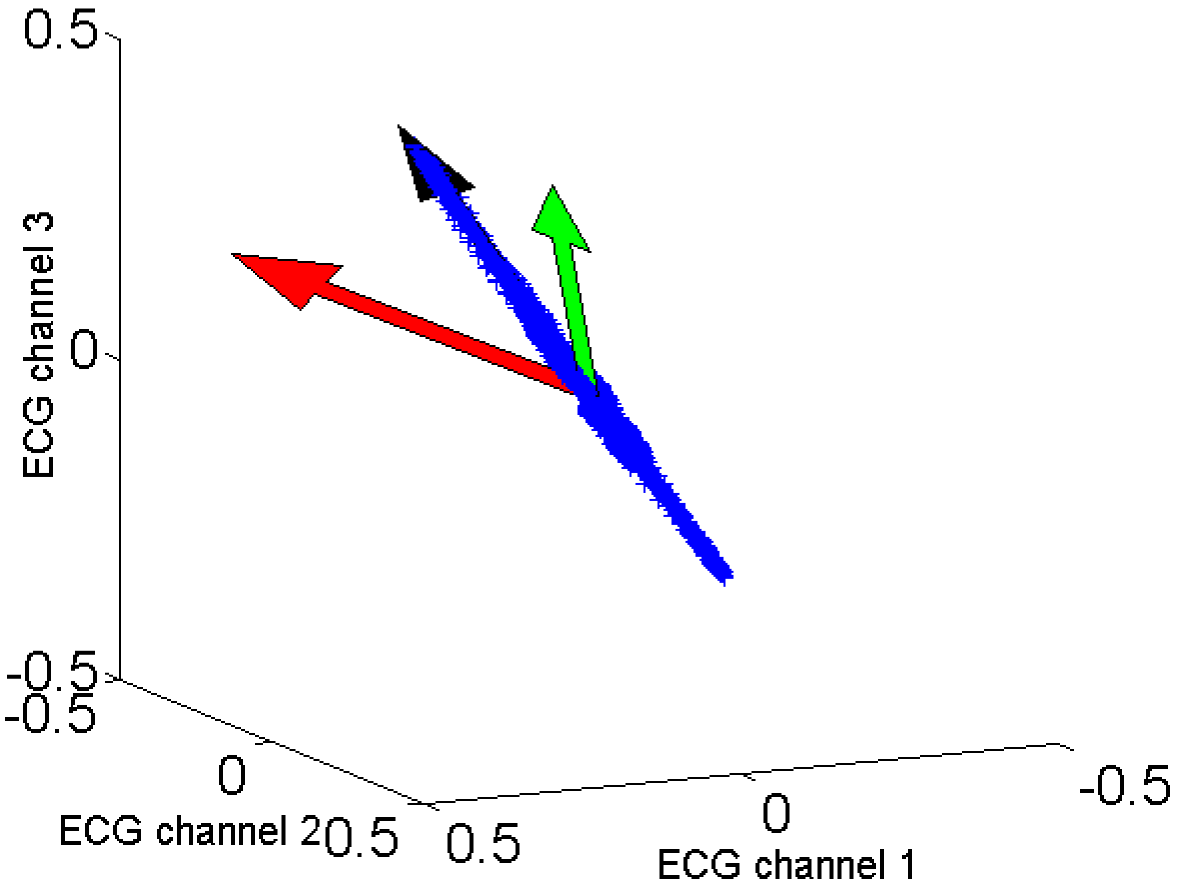
- Joint Approximation Diagonalization of Eigen-matrices (JADE)
  - Cardoso, Jean-François, and Antoine Souloumiac. "Blind beamforming for non-Gaussian signals." IEE proceedings F (radar and signal processing). Vol. 140. No. 6. IET Digital Library, 1993.
  - Cardoso, Jean-François. "High-order contrasts for independent component analysis." Neural computation 11.1 (1999): 157-192.
- FastICA
  - Hyvarinen, Aapo. "Fast and robust fixed-point algorithms for independent component analysis." IEEE transactions on Neural Networks 10.3 (1999): 626-634.
- ICA has been widely used in biosignals processing, in particular for ECG and EEG analysis.

AIMLab.

Quiz: axis

- Does ICA looks for orthogonal axes?

ICA **DOES NOT** require orthogonal axis



Quiz: whitening the data

- Does ICA preserves scaling?

$$\begin{cases} x_1(t) = a_{11}s_1 + a_{12}s_2 \\ x_2(t) = a_{21}s_1 + a_{22}s_2 \end{cases}$$

Both a and s are unknowns so any scalar multiplying one of the source s can be cancelled by dividing by the corresponding a by the same value.

ICA **DOES NOT** preserve scaling

Quiz: whitening the data

- Are uncorrelated and statistical independent variables the same thing?

Independant variables

$$p(y_1, y_2) = p_1(y_1)p_2(y_2)$$

Uncorrelated variables

$$E\{y_1 y_2\} - E\{y_1\}E\{y_2\} = 0$$

<u>Independence implies uncorrelated but not the opposite.</u>

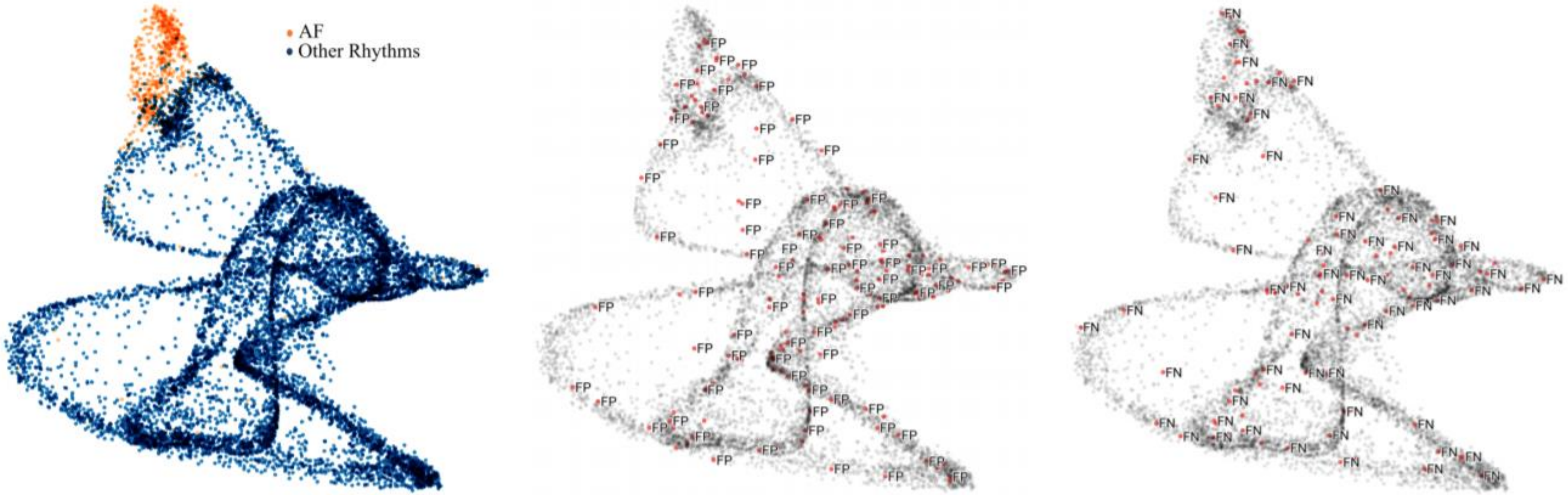
Beyond ICA

Beyond PCA and ICA

- PCA and ICA are commonly used algorithms in ML.
- However, often non-linear relationships between features exist and both standard PCA and ICA will not capture that.
- There are advanced techniques to tackle the limitation of PCA and ICA.
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is a probabilistic **non-linear technique** which is well suited for the visualization of high-dimensional datasets.
  - Application: e.g. image processing and genomic.
  - t-SNE is computationally expensive.
  - t-SNE is non-linear whereas PCA and ICA are linear.
- Uniform manifold approximation and projection UMAP…

Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605. https://lvdmaaten.github.io/tsne/
McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).
https://umap-learn.readthedocs.io/en/latest/

t-SNE



(a) Visualization of data belonging to *AF* and *Other Rhythm* class

(b) The False Positive (FP) cases highlighted

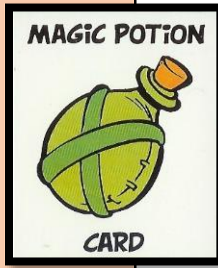
(c) The False Negative (FN) cases highlighted

Figure 5: Visualization of the features extracted from the deep learning pipeline and the time series covariates by performing t-SNE [37] based clustering. Clustering was performed on the testing dataset

Shashikumar, Supreeth P., et al. "Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018.

Take Home

- ICA look for **statistical independence**. It assumes a linear, instantaneous mixture of statistically independent sources (and not Gaussian).
- ICA exploits **spatial diversity**. Time structure are ignored.
- It looks to **maximize non-Gaussianity** of the sources it estimates.
- ICA is non-parametric.
- To solve the ICA problem we can use gradient descent on a cost function that we call the **contrast function**.



References

[1] Gari D. Clifford course note: http://www.mit.edu/~gari/teaching/6.555/SLIDES/BSShandouts.pdf

[2] Independent Component Analysis: Algorithms and Applications. Aapo Hyvärinen and Erkki Oja. URL: http://mlsp.cs.cmu.edu/courses/fall2012/lectures/ICA_Hyvarinen.pdf

[3] ICA for dummies. Online tutorial by Arnaud Delorme. URL: http://sccn.ucsd.edu/~arno/indexica.html

[4] Clifford, Gari D., and Francisco Azuaje. Advanced methods and tools for ECG data analysis. London: Artech house, 2006.

[5] Aapo Hyvarinen. tutorial on whitening. URL: http://cis.legacy.ics.tkk.fi/aapo/papers/IJCNN99_tutorialweb/node26.html

[6] Cardoso, J-F. "Blind signal separation: statistical principles." Proceedings of the IEEE 86.10 (1998): 2009-2025.

[7] Course notes: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf

[8] Course notes: http://cis.legacy.ics.tkk.fi/aapo/papers/NCS99web/node33.html

[9] Arnaud Delorme research page: http://arnauddelorme.com/ica_for_dummies/