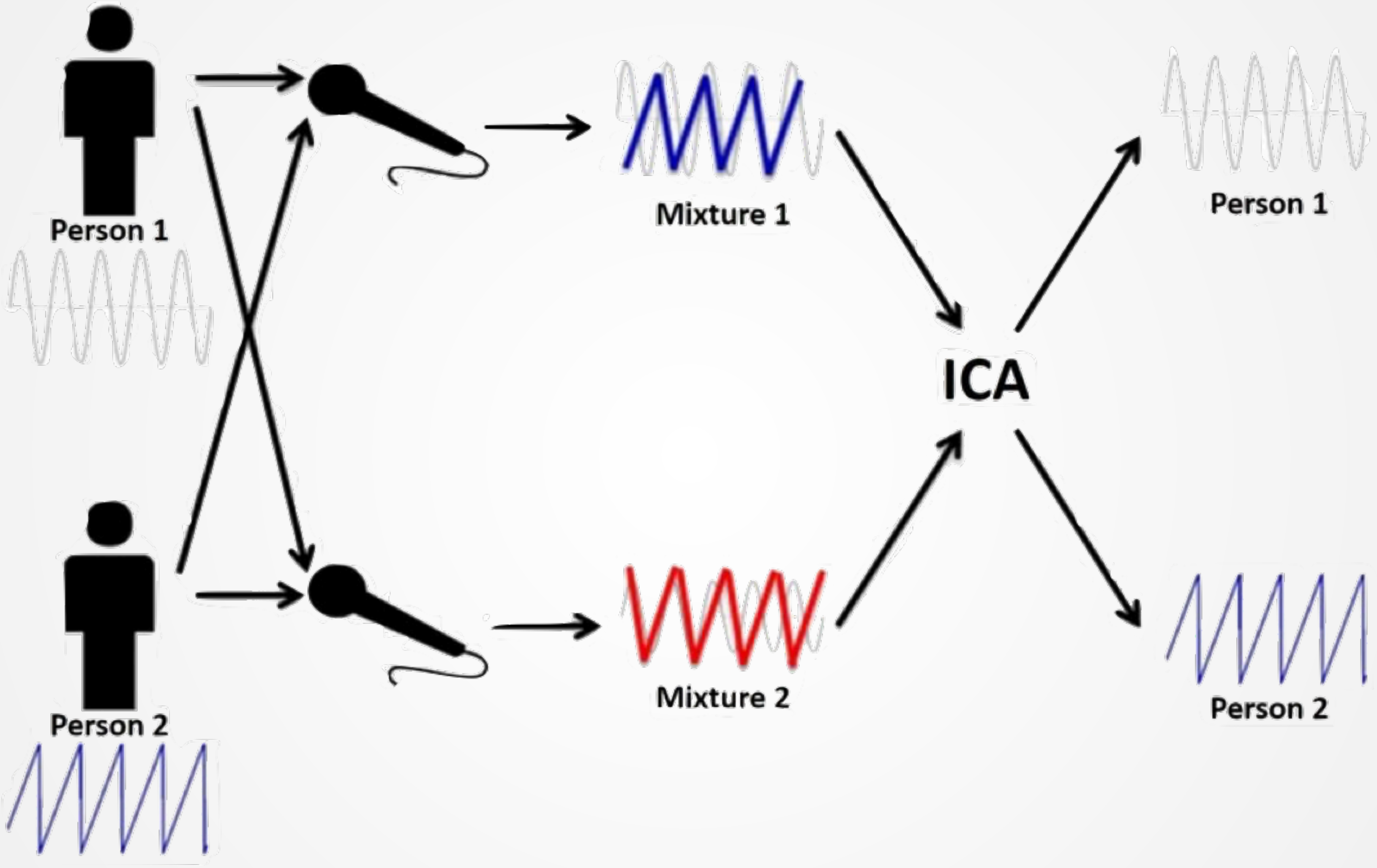
Independent Component Analysis

Jane Hsing-Chuan Hsieh
Wolfgang Karl Härdle

Ladislaus von Bortkiewicz Professor of Statistics
BRC Blockchain Research Center
International Research Training Group
Humboldt-Universität zu Berlin
lvb.wiwi.hu-berlin.de
www.case.hu-berlin.de
irtg1792.hu-berlin.de

Motivation

- Cocktail-party problem



- Blind source separation: Recover a set of source signals from a set of mixed signals

Cocktail-Party Problem

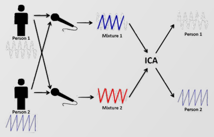
- Linear equation

$$x_1(t) = a_{11}y_1(t) + a_{12}y_2(t) \qquad (1)$$

$$x_2(t) = a_{21}y_1(t) + a_{22}y_2(t) \qquad (2)$$

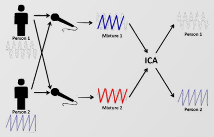
  - $x_1(t)$, $x_2(t)$ mixed signals, time index t
  - $y_1(t)$, $y_2(t)$ original signals

- Problem
  - Find a_{ij} i.e. solve (1)-(2)
  - crank out the original signals



Assumption

- Signal sources are statistically independent
  - i.e., signal sources are non-Gaussian (non-normal)
- (Optional) Number of ICs is equal to # of observed variables
  - A variant is with lower number of ICs allowed



Outline

1. Motivation ✓
2. Introduction of ICA
3. Preprocessing for ICA
4. Principles of ICA Estimation
5. FastICA Algorithm
6. ICA vs. FA vs. PCA
7. Applications
8. References

Definition

- General matrix form

$$\begin{bmatrix} y_{1t} \\ \vdots \\ y_{dt} \end{bmatrix} = \begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{d1} & \cdots & w_{dd} \end{bmatrix} \begin{bmatrix} x_{1t} \\ \vdots \\ x_{dt} \end{bmatrix} \tag{3}$$

equivalently
$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) = [\mathbf{w}_1, \ldots, \mathbf{w}_d]^\top \mathbf{x}(t) \tag{4}$$

$$\mathbf{x}(t) = \mathbf{A}\mathbf{y}(t) \tag{5}$$

where

$\mathbf{y}(t)$: *d*-dim „independent components (ICs)“ or „source signals“

$\mathbf{x}(t)$: *d*-dim observations; w.l.o.g., $\mathsf{E}\{\mathbf{x}(t)\} = 0$

$\mathbf{W}$: nonsingular linear transformation matrix, s.t. $\mathbf{W}^{-1} = \mathbf{A}$.

$\mathbf{A}$: matrix of $[a_{ij}]$ is a constant parameter called “mixing matrix”

From now on, time index *t* is dropped for specification simplicity

Ambiguities of ICA

Scale identification

Cannot determine the variances of the ICs

- any scalar multiplier in one of the sources cancels out by dividing the corresponding column of $\mathbf{A}$ by the same scaler

$$x_i = \sum_{j=1}^{d} a_{ij} y_j = \sum_{j=1}^{d} (c_j^{-1} a_{ij})(c_j y_j)$$

- hence fix the magnitudes (scale) of the ICs to unit variance

From now on assume this

$$\mathrm{Var}\{y_j\} = \mathrm{E}\{y_j^2\} = 1 \qquad (6)$$

- still leaves the ambiguity of the sign; i.e., multiplying IC by -1 won’t affect the model results

Ambiguities of ICA

⊡ **Order identification**

Like in PCA cannot determine the order of the IC directions

$$\mathbf{x} = \mathbf{A}\mathbf{y} = (\mathbf{A}\mathbf{P}^{-1})(\mathbf{P}\mathbf{y})$$

- ▶ where $\mathbf{P}$ is a permutation matrix and $\mathbf{P}\mathbf{y}$ are the original ICs but in a different order
- ▶ i.e., any permutation transformation of independent sources are still independent and ICA can not distinguish between them

Preprocessing for ICA

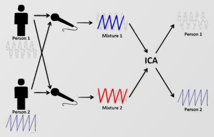
- ⊡ Data preprocessing before ICA, to make its problem of estimation simpler and better conditioned

- ⊡ **Centering**
  - ▶ center $\mathbf{x}$ into $\tilde{\mathbf{x}}$ by subtracting its mean vector $\mathbf{m} = \mathrm{E}\{\mathbf{x}\}$, s.t. $\mathrm{E}\{\tilde{\mathbf{x}}\} = 0$

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{m}$$

  - ▶ which implies $\mathrm{E}\{\mathbf{y}\} = 0$ as well

$$\mathrm{E}\{\mathbf{y}\} = \mathrm{E}\{\tilde{\mathbf{W}}\tilde{\mathbf{x}}\} = \tilde{\mathbf{W}}\mathrm{E}\{\tilde{\mathbf{x}}\} = \mathbf{0}$$



Preprocessing for ICA

⊡ Whitening

- ▶ transform centered $\mathbf{x}$ into $\tilde{\mathbf{x}}$ through eigenvalue decomposition (EVD), s.t. $\mathrm{E}\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top\} = \mathbf{I}$

$$\tilde{\mathbf{x}} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{x} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{A}\mathbf{y} = \tilde{\mathbf{A}}\mathbf{y}$$

new mixing matrix

where

$\mathbf{E}$: orthogonal eigenvector matrix of $\mathrm{Cov}(\mathbf{x})$

$\mathbf{D}$: diagonal eigenvalue matrices of $\mathrm{Cov}(\mathbf{x})$

- ▶ $\tilde{\mathbf{A}}$ is hence orthogonal, which reduces its number of parameters to be estimated, from d^2 to $d(d-1)/2$ *df*

$$\mathrm{E}\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top\} = \tilde{\mathbf{A}}\mathrm{E}\{\mathbf{y}\mathbf{y}^\top\}\tilde{\mathbf{A}}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top = \mathbf{I}$$

$\mathrm{E}\{\mathbf{y}\mathbf{y}^\top\} = \mathbf{I}$,since IC is independent and from (6)

What is Independence?

⊡ Definition of independence

- ▶ variables y_1 and y_2 are said to be independent if information on y_1 does not give any information on y_2, and vice versa
- ▶ technically, y_1 and y_2 are independent iff the joint pdf $p(y_1, y_2)$ is factorizable by their marginal pdf's

$$p(y_1, y_2) = p_1(y_1)p_2(y_2) \qquad (7)$$

- ▶ from (7), given any function h_1, h_2, we always have

$$\mathrm{E}\{h_1(y_1) * h_2(y_2)\} = \mathrm{E}\{h_1(y_1)\}\ \mathrm{E}\{h_2(y_2)\} \qquad (8)$$

⊡ Uncorrelatedness

- ▶ y_1 and y_2 are said to be uncorrelated, if their covariance is zero

$$\mathrm{E}\{y_1 * y_2\} - \mathrm{E}\{y_1\}\ \mathrm{E}\{y_2\} = 0 \qquad (9)$$

- ▶ (8) $\succ$ (9), but (9) $\not\succ$ (8)

Uncorrelatedness does not imply independence

- Generate ICA Example
  - given **independent** signals $\mathbf{y} = [y_1, y_2, y_3]^\top$ ($\succ$ $\mathrm{E}\{\mathbf{y}\} = \mathbf{0}$, $\mathrm{E}\{\mathbf{y}\mathbf{y}^\top\} = \mathbf{I}$)
  - suppose

$$\mathbf{A} = \mathbf{W}^{-1} = \begin{bmatrix} 1.31 & 0.14 & 0.18 \\ -0.42 & -1.26 & -1.25 \\ -0.03 & 0.41 & -0.49 \end{bmatrix} 10^{-2}$$

  - from (5), derive $\mathbf{x} = \mathbf{A}\mathbf{y}$ (i.e., $\mathbf{y} = \boxed{\mathbf{W}}\mathbf{x}$)

- Mahalanobis transformation to $\mathbf{x}$
  - whiten $\mathbf{x} \succ \mathbf{y}'$ below, s.t. y_i' and y_j' become **uncorrelated** for all *i*, *j*

$$\mathbf{y}' = \boxed{\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-\frac{1}{2}}}\mathbf{x} \quad \succ \quad \mathrm{E}\{y'\} = \mathbf{0}, \quad \mathrm{E}\{\mathbf{y}'\mathbf{y}'^\top\} = \mathbf{I}$$

Uncorrelatedness does not imply independence

- However,

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-\frac{1}{2}} = \begin{bmatrix} 1.39 & 0.20 & -0.01 \\ 0.20 & 0.71 & -0.22 \\ -0.01 & -0.22 & 1.92 \end{bmatrix} 10^2$$

$$\neq \quad \mathbf{W} = \begin{bmatrix} 0.80 & 0.10 & 0.04 \\ -0.12 & -0.45 & 1.10 \\ -0.15 & -0.38 & -1.12 \end{bmatrix} 10^2$$

- Cross-cumulants
  - signals from ICA transformation are independent, whereas not the case from Mahalanobis transformation

Transformation	Mahalanobis	ICA
E$\{y_1 y_3^2\}$	0.0643	-0.0004
E$\{y_2^2 y_3\}$	-0.0050	-0.0029
E$\{y_1^3 y_2\}$	0.0553	0.0062
E$\{y_1^2 y_2 y_3\}$	0.0264	-0.0253

ICs are necessarily non-Gaussian

- ICs must be non-Gaussian for an ICA!
- Counterexample
  - consider any random orthogonal mixing matrix $\mathbf{A}$, and two ICs y_1 and y_2 which are **Gaussian**, then

$$f(y_1, y_2) = |2\pi\mathbf{I}|^{-\frac{1}{2}} \exp(-\frac{y_1^2 + y_2^2}{2}) = \frac{1}{2\pi} \exp(-\frac{\|\mathbf{y}\|^2}{2})$$

where $\|\mathbf{y}\|$ is the norm of the vector (y_1, y_2)

Exactly same distribution!

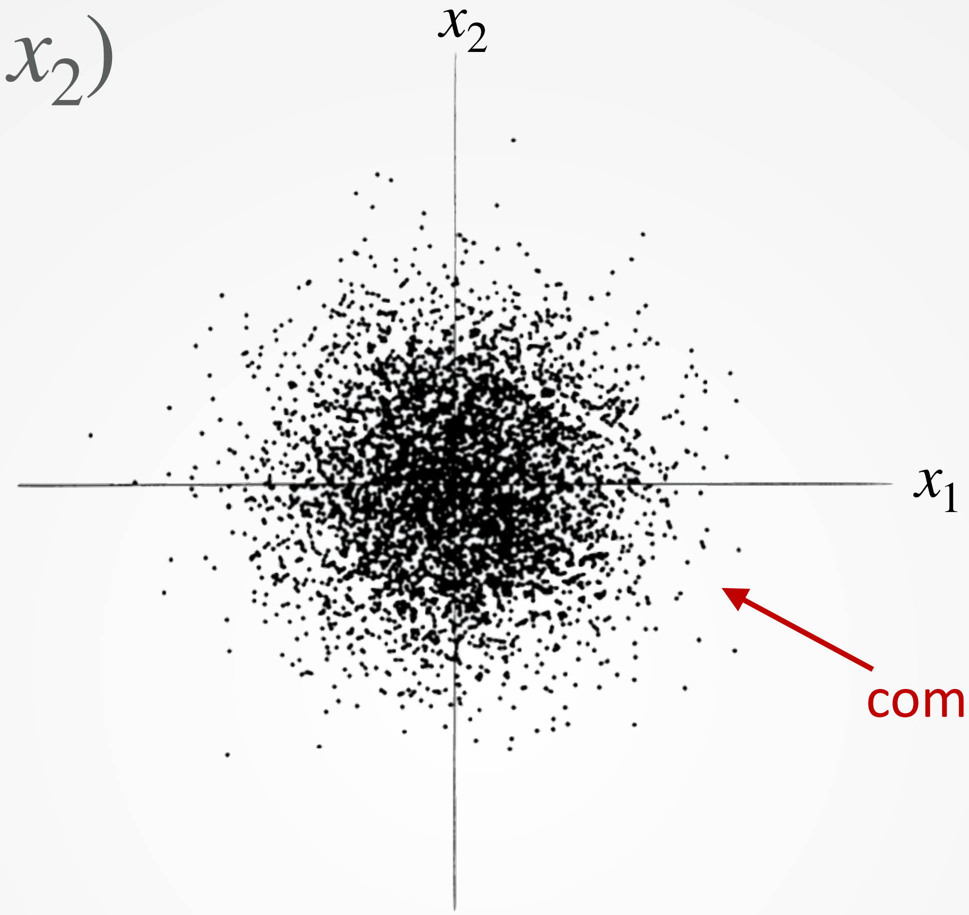
  - then x_1 and x_2 are **Gaussian**, uncorrelated, and of unit variance

$$f(x_1, x_2) = |2\pi\mathbf{I}|^{-\frac{1}{2}} \exp(-\frac{\|\mathbf{Wx}\|^2}{2})\,|\det \mathbf{W}| = \frac{1}{2\pi} \exp(-\frac{\|\mathbf{x}\|^2}{2})$$

$\because$ $\mathbf{A}$ is orthogonal $\succ$ $|\det \mathbf{W}| = |\det \mathbf{A}^{-1}| = |\det \mathbf{A}^{\top}| = 1$

ICs are necessarily non-Gaussian

- Joint density $f(x_1, x_2)$



completely symmetric!

  - it thus does not contain any information on the directions of the columns of mixing matrix $\mathbf{A}$, making it unable to be estimated
  - in other words, $\mathbf{A}$ is not identifiable for Gaussian independent components

Non-Gaussian is independent

- ⊡ Central Limit Theory (CLT)

  sum of independent random variables will have a distribution that is close(r) to Gaussian

  - ▶ any mixture of components will be more Gaussian than any of the components themselves

- ⊡ Use CLT to approximate $\mathbf{w}$ (i.e., rows of $\mathbf{A}^{-1}$) from data vector $\mathbf{x}$ (Equ. (4)-(5))

  - ▶ define $\mathbf{z} = \mathbf{A}^\top \mathbf{w}$, then we have one of the ICs:

    $$\hat{y} = \mathbf{w}^\top \mathbf{x} = \mathbf{w}^\top \mathbf{A} \mathbf{y} = \mathbf{z}^\top \mathbf{y} \longleftarrow \text{more Gaussian than any of } y_j\text{'s by CLT}$$

  - ▶ $\mathbf{z}^\top \mathbf{y}$ becomes least Gaussian when it in fact equals one of $y_j \longleftarrow$ i.e., $\hat{y}$
  - ▶ therefore, $\mathbf{w}$ can be estimated by maximizing the non-Gaussianity of $\mathbf{w}^\top \mathbf{x} = \mathbf{z}^\top \mathbf{y}$

Non-Gaussian measures

1. Kurtosis: cumulant-based estimator

$$\text{kurt}\{y\} = \text{E}\{y^4\} - 3(\text{E}\{y^2\}) \qquad (10)$$

$$= \text{E}\{y^4\} - 3 \text{ from (6)}$$

- zero for a Gaussian variable, and non-zero for most non-Gaussian random variables

- non-Gaussianity is typically measured by maximizing $|\,\text{kurt}\{y\}\,|$ or $\text{kurt}\{y\}^2$

- cumulant-based estimator is classic but sensitive to outliers, thus not a robust measure of non-Gaussianity

Non-Gaussian measures

2. Negentropy:

$$J(y) = H(y_{gauss}) - H(y) \qquad (11)$$

- ▶ differential entropy H measures degree of randomness or unpredictability of the variable

$$H(y) = -\int f(y) \log f(y) dy \qquad (12)$$

- ▶ among all random variables of equal variance, $H(y)$ is largest if y is Gaussian (i.e., $H(y_{gauss})$) ⟵ $H(\mathrm{N}\{0,1\})$ from (6)
- ▶ non-Gaussianity is thus typically measured by

$$\operatorname{argmax}\{J(y)\} = \operatorname{argmin}\{H(y)\}$$

- ▶ statistically better, but difficult to compute ⟵ requires the knowledge of $f(y)$

  ≻ a practical compromise – approximations of Negentropy

Non-Gaussian measures

3. Mutual information: a natural measure of the dependence between random variables

$$I(\mathbf{y}) = \sum_{j=1}^{d} H(y_j) - H(\mathbf{y}) \qquad (13)$$

from (4), and $\det(\mathbf{W}) = 1$ for orthogonal $\mathbf{W}$

$$= \sum_{j=1}^{d} H(y_j) - (H(\mathbf{x}) + \log|\det(\mathbf{W})|) \propto \sum_{j=1}^{d} H(y_j) \propto -\sum_{j=1}^{d} J(y_j)$$

- maximizing the sum of Negentropy of y_j's is equivalent to minimization of mutual information
- non-Gaussianity is thus measured by minimization of $I(y)$; furthermore,

$$\min \sum_{j=1}^{d} H(y_j) \geq \sum_{j=1}^{d} \min H(y_j)$$

$$\succ \hat{\mathbf{w}}_j = \operatorname{argmin} H(y_j) = \operatorname{argmax} J(y_j)$$

So IC and be found by optimizing univariate (neg)entropy from (12) - (13)!

How to approximate univariate entropy – $H(y)$?

- Maximum entropy method (Hyvärinen, 1998)
  - Assume given contrast functions $G_i(y)$ has fixed expectation c_i on $f(y)$

$$\mathrm{E}\{G_i(y)\} = \int G_i(y)f(y)dy = c_i, \qquad i = 1,\ldots,s \qquad (14)$$

  - Problem: without knowledge of $f(y)$, there exists an infinite number of pdfs compatible with (14), but with very different differential entropies from (12)
  - Simple solution: maximizing entropy $H(y)$ constrained by (14), as an approximation of the entropy of y.
    - ≻ Not minimization, since $H(y)$ reaches $-\infty$ in the limit where y takes only a finite number of values

How to approximate univariate entropy – $H(y)$?

- □ Approximated pdf from maximum entropy method satisfying (14) (Cover & Thomas, 2002)

$$f_0(y; \mathbf{b}) = \mathbf{B} \exp \sum b_i G_i(y) \qquad (15)$$

- ▶ constants $\mathbf{B}$ and b_i can be determined from c_i, using the constraints (14) and that $\int f_0(y)dy = 1$ ← Hard to solve $\mathbf{B}$ and b_i !

- □ To simplify calculation, we construct a 1st-order density expansion in the vicinity of the standard Normal – $\varphi(y)$; additionally,
  - ▶ add 2 additional constraints to (14) for standardization

$$G_{s+1}(y) = y, c_{s+1} = 0; \quad G_{s+2}(y) = y^2, c_{s+2} = 1 \qquad (16)$$

  - ▶ Transform G_i's into an orthonormal system, and orthogonal to all 2nd-order polynomials (Hyvärinen, 1998) (17)

How to approximate univariate entropy – $H(y)$?

- Approximated pdf from maximum entropy method satisfying (14), (16), (17):

$$\hat{f}_y = \varphi(y)\{1 + \sum_{i=1}^{s} c_i G_i(y)\} \qquad (18)$$

  - in practice, each c_i is estimated as the sample average of $G_i(y)$, in order to estimate $\hat{f}_y$

- Hence we can approximate the negentropy from (18):

$$H(y) \approx -\int \hat{f}_y(u) \log \hat{f}_y(u) du \approx H(y_{Gauss}) - \frac{1}{2} \sum_{i=1}^{s} c_i^2 \qquad (19)$$

$$J(y) = H(y_{Gauss}) - H(y) = \frac{1}{2} \sum_{i=1}^{s} c_i^2 \qquad (20)$$

  - Proof in Appendix of (Hyvärinen et al., 2001)

How to approximate univariate entropy – $H(y)$?

- Choose contrast functions G_i

1. E$\{G_i(y)\}$ should be easily computable and not sensitive to outliers
2. $G_i(y)$ should not grow faster than quadratically to ensure that $f_0(y)$ in (15) is integrable
3. $G_i(y)$ should capture distributional features of $\log\{f_y(y)\}$

How to approximate univariate entropy – $H(y)$?

- Choose contrast functions G_i – special case
  - Two most important features to measure non-Gaussianity (s=2)
    1. Asymmetry - G_1 as odd function
    2. Tail behavior - G_2 as even function
  - Hence the approximation in (20) can be simplified to

$$J(y) \approx \frac{1}{2}\sum_{i=1}^{s=2} c_i^2 \approx \sum_{i=1}^{s=2} k_i[\,\mathrm{E}\{G_i(y)\} - \mathrm{E}\{G_i(y_{Gauss})\}\,]^2$$

$$\approx k_1 \mathrm{E}\{G_1(y)\}^2 + k_2[\,\mathrm{E}\{G_2(y)\} - \mathrm{E}\{G_2(y_{Gauss})\}\,]^2 \qquad (21)$$

where k_1, k_2: positive constants

  - Proof in Appendix of (Hyvärinen et al., 2001)

Example: Negentropy approximation

- ⊡ Approximation a:

$$J(y) \approx k_1 \mathrm{E}\{y\exp(-y^2/2)\}^2 + k_2^a[\,\mathrm{E}\{\exp(-y^2/2)\} - \sqrt{1/2}\,]^2$$

$$G_1(y) = y\exp(-y^2/2)$$

$$G_2^a(y) = \exp(-y^2/2)$$

- ⊡ Approximation b:

$$J(y) \approx k_1 \mathrm{E}\{y\exp(-y^2/2)\}^2 + k_2^b[\,\mathrm{E}\{\,|y|\,\} - \sqrt{2/\pi}\,]^2$$

$$G_1(y) = y\exp(-y^2/2)$$

$$G_2^b(y) = |y|$$

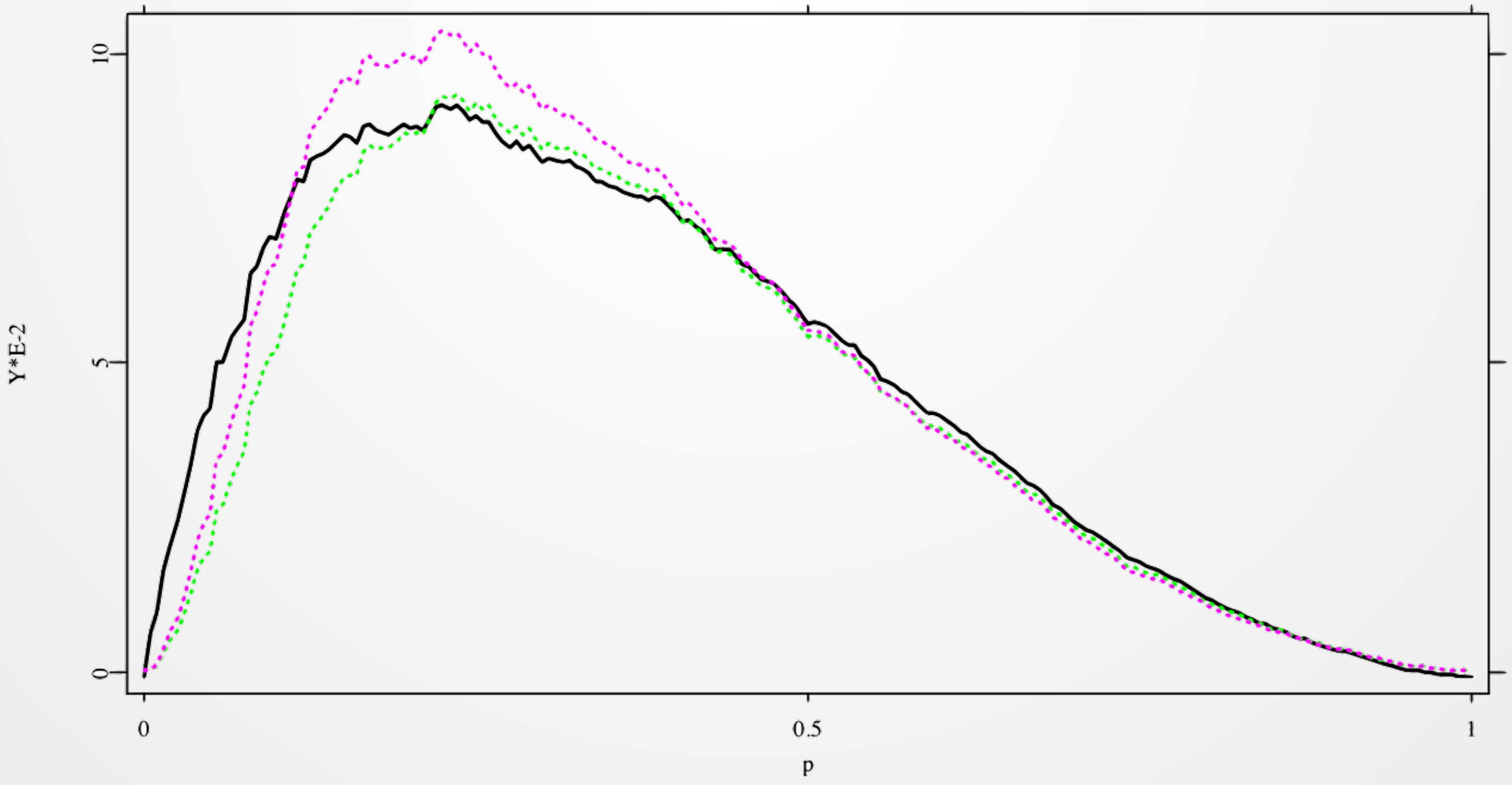
- ▶ $k_1 = 36/(8\sqrt{3} - 9)$
  $k_2^a = 1/(2 - 6/\pi)$, $k_2^b = 24/(16\sqrt{3} - 27/\pi)$

Example: Negentropy approximation

- Example: Negentropy approximation
  - Comparison of the true negentropy (black) and its approximations (a: pink, b: green) of simulated Gaussian mixture variable:

$$y \sim p\,\mathrm{N}(0,1) + (1-p)\mathrm{N}(1,4), \quad p \in [0,1]$$

Negentropy comparison



Optimization algorithm

⊡ Given any $y = \mathbf{w}^\top \mathbf{x}$ and chosen contrast function G_i's, We can now approximate the corresponding negentropy from (21)

⊡ Next step: find the optimized IC (aka $\hat{y}$ via estimating $\hat{\mathbf{w}}$) and $\hat{G}_i$'s with the highest negentropy

⊡ Optimization algorithms

1. Fixed-point algorithm: FastICA (Hyvärinen et al., 2001) ✓
2. Gradient methods: infomax (Lee et al., 1999)

Negentropy approximations and FastICA

- ⊡ In the case where we use only one non-quadratic function G (21), the approximation becomes

$$J(y) \propto [\,\mathrm{E}\{G(y)\} - \mathrm{E}\{G(y_{Gauss})\}\,]^2 \qquad (22)$$

- ⊡ The following choices of G have proved very useful

$$G^a(y) = \frac{1}{a_1} \log\cosh(a_1 y), \quad G^b(y) = -\exp(-y^2/2)$$

  - ▶ $1 \leq a_1 \leq 2$: suitable constant

  - ▶ The respective first derivatives of G are

$$g^a(y) = \tanh(a_1 y), \quad g^b(y) = y\exp(-y^2/2)$$

FastICA

- FastICA finds a direction, i.e. a unit vector $\mathbf{w}$ s.t. the projection $y = \mathbf{w}^\top \mathbf{x}$ maximizes non-Gaussianity from (22)

- **Objective function**

$$\max_{\mathbf{w},\, \|\mathbf{w}\|^2=1} \mathrm{E}\{G(y)\} = \max_{\mathbf{w},\, \|\mathbf{w}\|^2=1} \mathrm{E}\{G(\mathbf{w}^\top \mathbf{x})\}$$

$\because$ $\mathrm{E}\{(\mathbf{w}^\top \mathbf{x})^2\} = \|\mathbf{w}\|^2 = 1$ from (6)

- according to Kuhn-Tucker conditions, points of optima are at

$$\mathrm{E}\{\mathbf{x} g(\mathbf{w}^\top \mathbf{x})\} - \beta \mathbf{w} = 0$$

Lagrange multiplier

- approximative Newton iteration (Hyvärinen & Oja, 1999) shows

$$\mathbf{w}^+ = \mathbf{w} - [\,\mathrm{E}\{\mathbf{x} g(\mathbf{w}^\top \mathbf{x})\} - \beta \mathbf{w}\,] / [\,\mathrm{E}\{g'(\mathbf{w}^\top \mathbf{x})\} - \beta\,]$$

simplified by multiplying both sides by $[\,\mathrm{E}\{g'(\mathbf{w}^\top \mathbf{x})\} - \beta\,]$

$$\succ\ \mathbf{w}^+ = \mathrm{E}\{\mathbf{x} g(\mathbf{w}^\top \mathbf{x})\} - \mathrm{E}\{g'(\mathbf{w}^\top \mathbf{x})\}\mathbf{w} \qquad (23)$$

FastICA

⊡ Algorithm

1. Initialize index $j = 1$, and set of indices $K = \{j\}$
2. Choose an initial vector $\mathbf{w}_j$ of unit norm, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d]^\top$
3. Update $\mathbf{w}_j \rightarrow \mathbf{w}_j^+$ according to (23)

   In practice, the sample mean is used to calculate *E[·]*
4. Orthogonalization (Gram-Schmidt-liked decorrelation):

$$\mathbf{w}_j^+ = \mathbf{w}_j^+ - \sum_{k \in K,\, k \neq j} (\mathbf{w}_j^{+\top} \mathbf{w}_k)\mathbf{w}_k$$

5. Normalization:

$$\mathbf{w}_j^+ = \mathbf{w}_j^+ / \|\mathbf{w}_j^+\|$$

6. If not converged (i.e. $\|\mathbf{w}_j^+ - \mathbf{w}_j\| \neq 0$), then

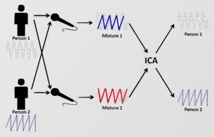
   set $\mathbf{w}_j = \mathbf{w}_j^+$, and go back to 3
7. Set $j = j + 1$, $K = K \cup \{j\}$. For $j \leq d$, go back to step 2

Independent Component Analysis (ICA) – Review

- **Scenarios**: Blind Source Separation

- **Objective**: find direction $\mathbf{y} = \mathbf{w}^\top \mathbf{x} = \sum w_j x_j$ of statistically independence

- **Properties**
  - Components $\mathbf{y}$ are assumed statistically independent and non-Gaussian
  - Importances of $\mathbf{y}$ can not be ordered
  - ICA is a higher-order statistical method due to stronger independence assumption (8)



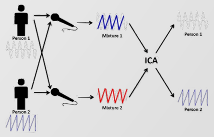
Principle Component Analysis (PCA)

- **Scenarios**: feature extraction, data compression
  - to reduce data redundancy

- **Objective**: find directions $y = \sum w_j x_j$ of maximum variance

- **Properties**
  - Components $\mathbf{y}$ are assumed uncorrelated, without explicit assumption of its pdf
  - Importances of $\mathbf{y}$ can be ordered by their variances (aka eigenvalues)
  - PCA is a 2nd-order statistical method due to uncorrelatedness assumption (9)

    $\succ$ only $d(d+1)/2$ parameters of covariances $\mathbf{cov}(x_j, x_k)$ needed for estimation)



Factor Analysis (FA)

- **Scenarios**: psychometrics, personality theories;
  - to find meaningful (lower-dim.) factors $\mathbf{y}$ that explain observants $\mathbf{x}$

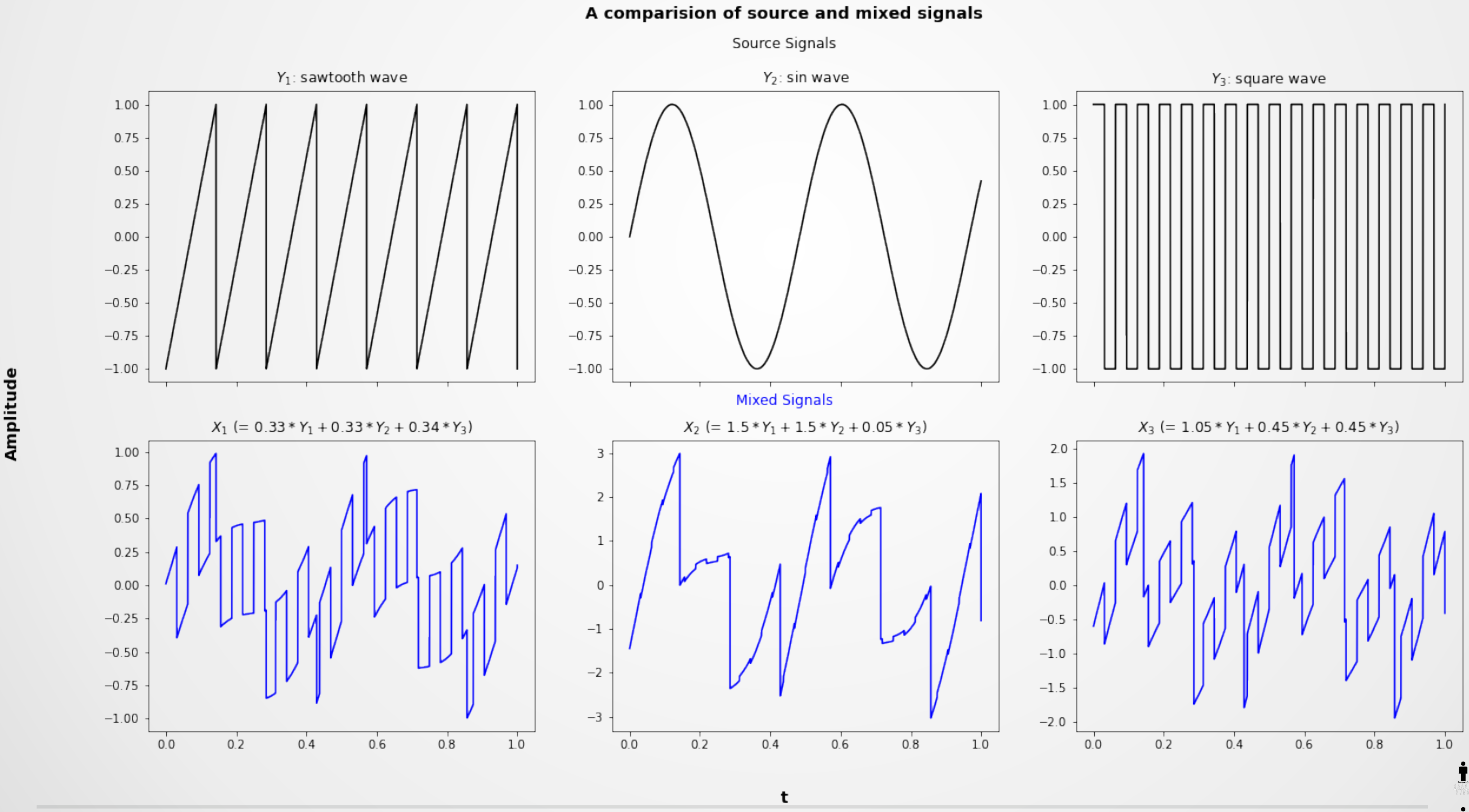
$$\mathbf{x} = \mathbf{A}\mathbf{y} + \varepsilon$$

Notice additional noise term than (5)

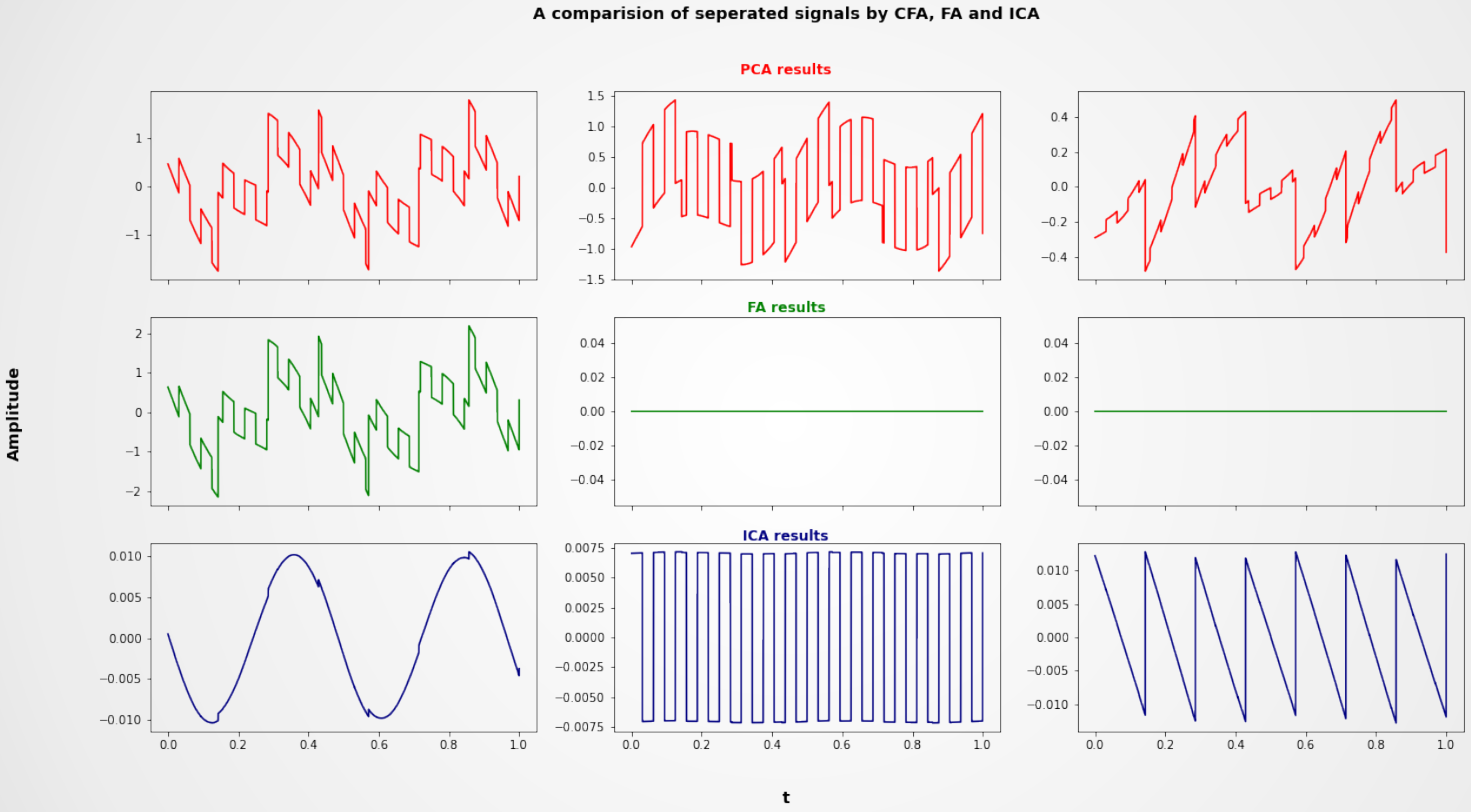
- **Objective**: find pattern $\mathbf{A}$ as simple as possible;
  - i.e., each $x = \sum a_j y_j$ loads strongly only on one factor ($a_j \neq 0$), and much weaker on the others ($a_k \approx 0, \forall k \neq j$)
- **Properties**
  - Factors $\mathbf{y}$ are uncorrelated and Gaussian (thus independent for Gaussian case)
  - $\mathbf{A}$ is not unidentifiable due to Gaussianity, making further formulation of its simple structure necessary
  - FA is also a 2nd-order statistical method (same as PCA)

E.g. 1. Generated Artificial Data (for demonstration)

- We generate 3 independent signals ($\mathbf{Y}_1 \sim \mathbf{Y}_3$), which are mixed together with different ratios to form mixed signals ($\mathbf{X}_1 \sim \mathbf{X}_3$)



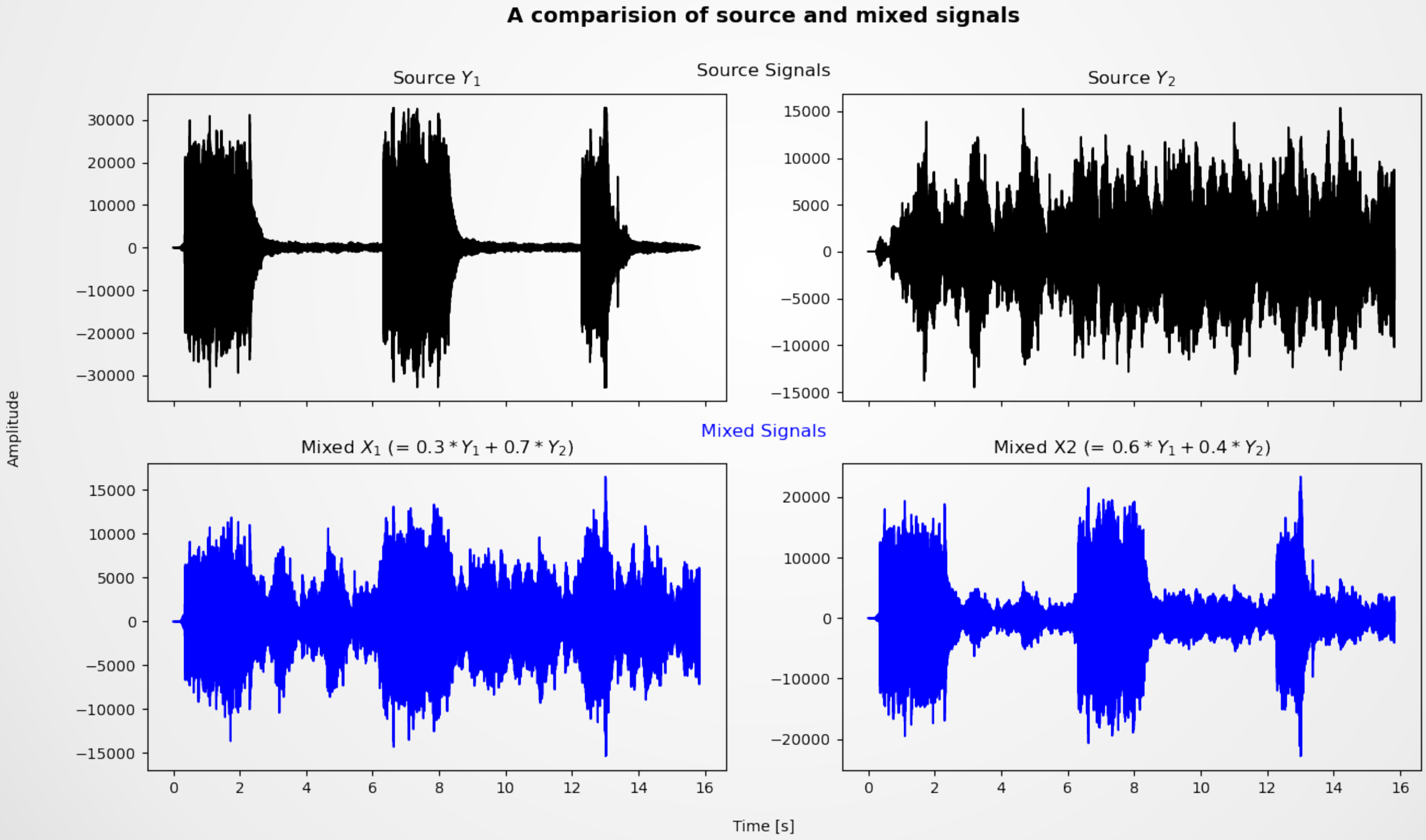
E.g. 1. Generated Artificial Data (for demonstration)



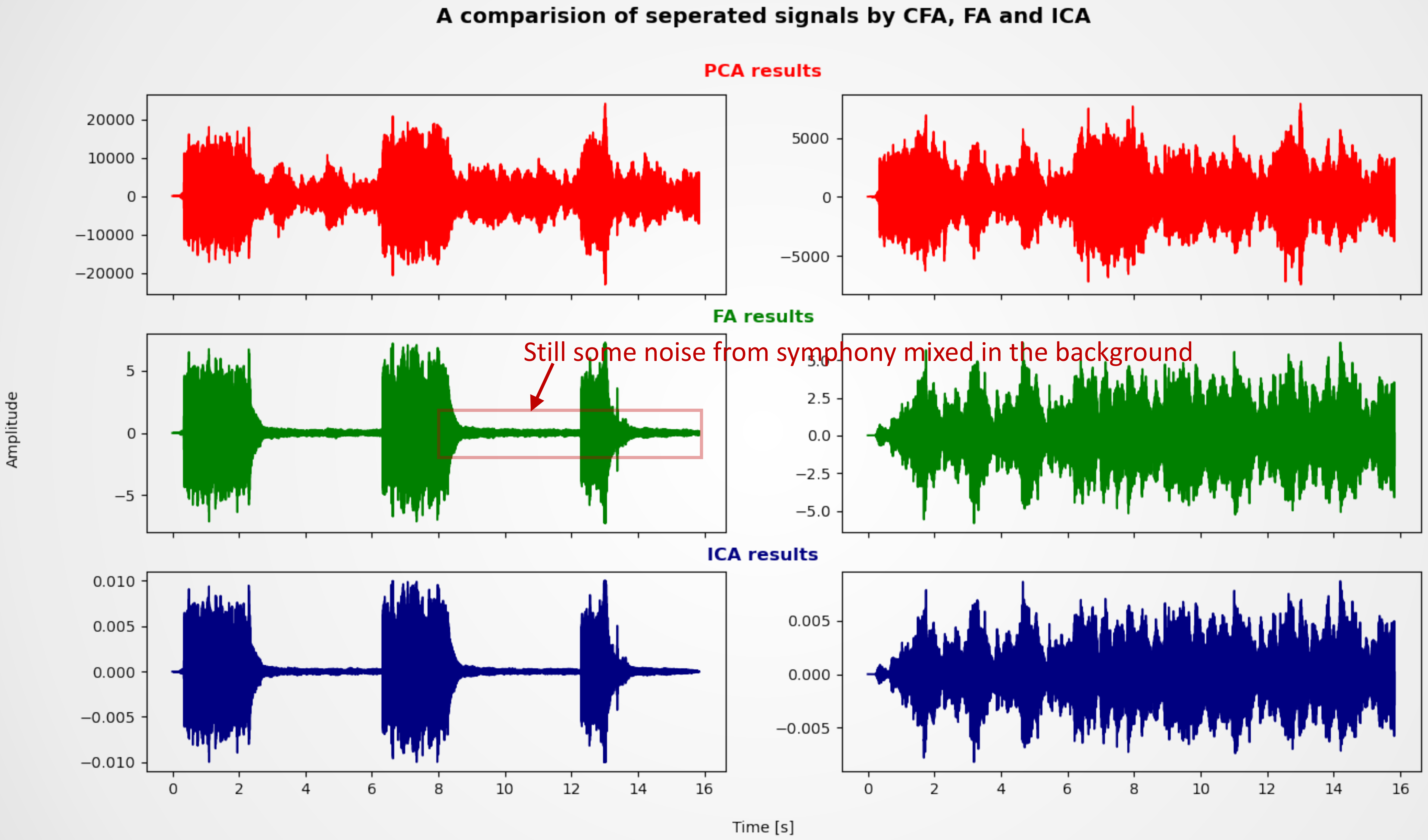
- While PCA and FA both fail at separating independent (similar to the case of Mahalanobis transformation), ICA, however, succeeds in better approximation.

E.g. 2. Cocktail Party Problem (or Blind Source Separation)

- There're 2 sources of sounds (bell rings & symphony), which are received by 2 microphones (and the sounds are mixed together with different ratios)

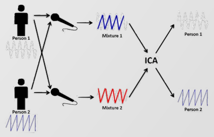


E.g. 2. Cocktail Party Problem (or Blind Source Separation)



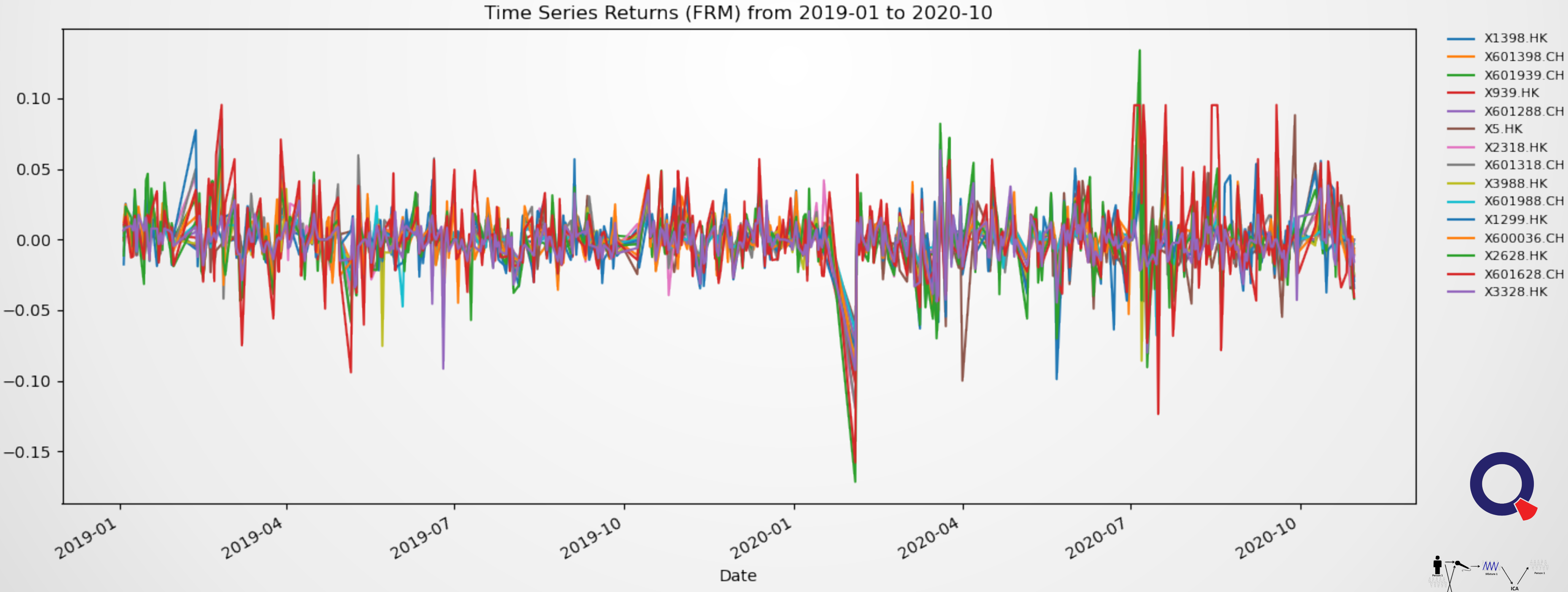
Still some noise from symphony mixed in the background

- ⊡ While PCA cannot separate source signals, FA do a better job with minor noises. However, ICA can totally separate the signals.

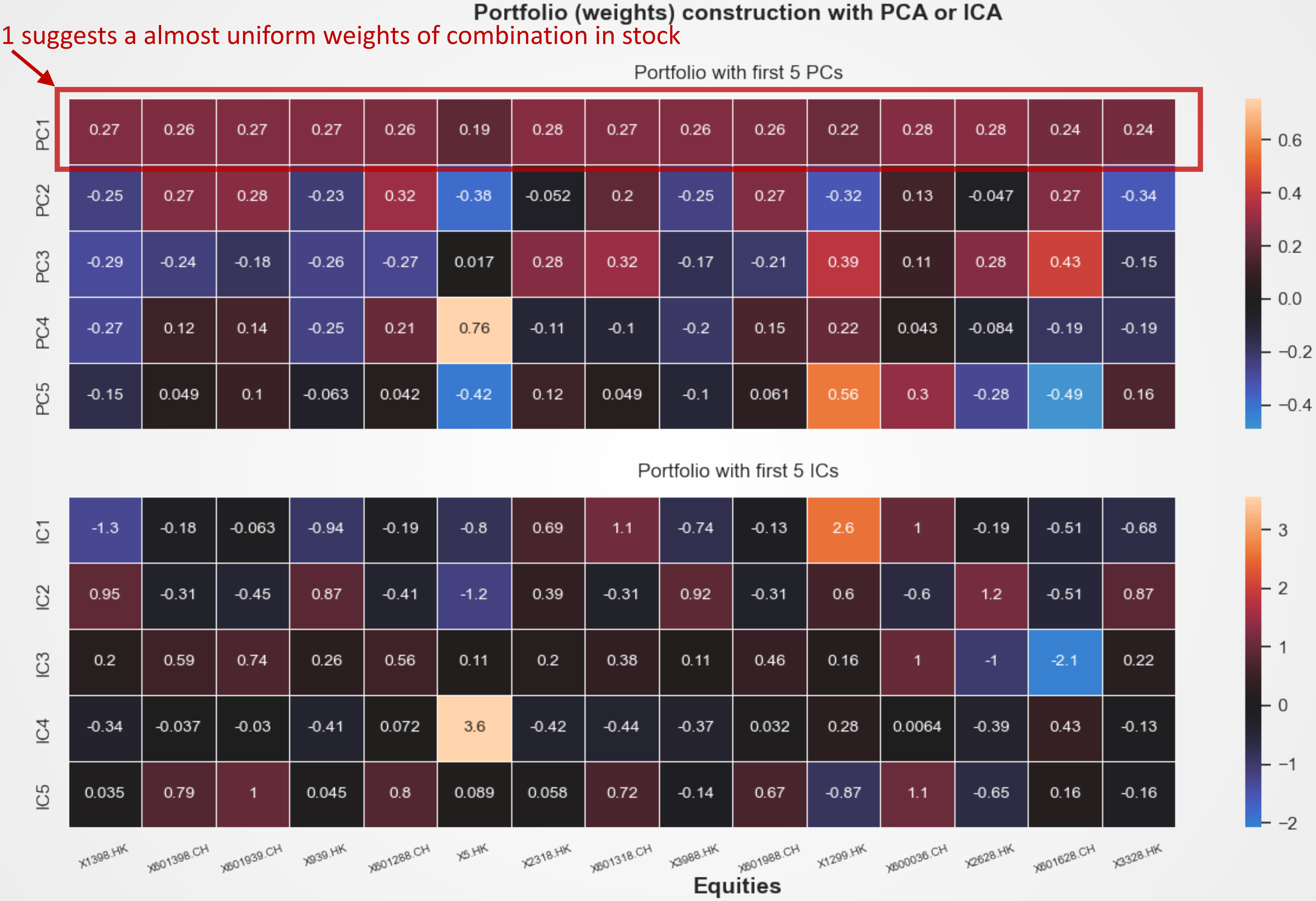


E.g. 3. Daily Stock Returns (15 Equities in Asia Market)

- Purpose: Optimal portfolio construction for 15 stocks in Asia Market.
  - Build different strategies for combination of stock investments (i.e., estimation of each component (by PCA or ICA) helps with deciding the weights of investment for each selected stock) (Rosenzweig, 2020).



E.g. 3. Daily Stock Returns (15 Equities in Asia Market)



Portfolio (weights) construction with PCA or ICA

Strategy on PC1 suggests a almost uniform weights of combination in stock investment

Portfolio with first 5 PCs

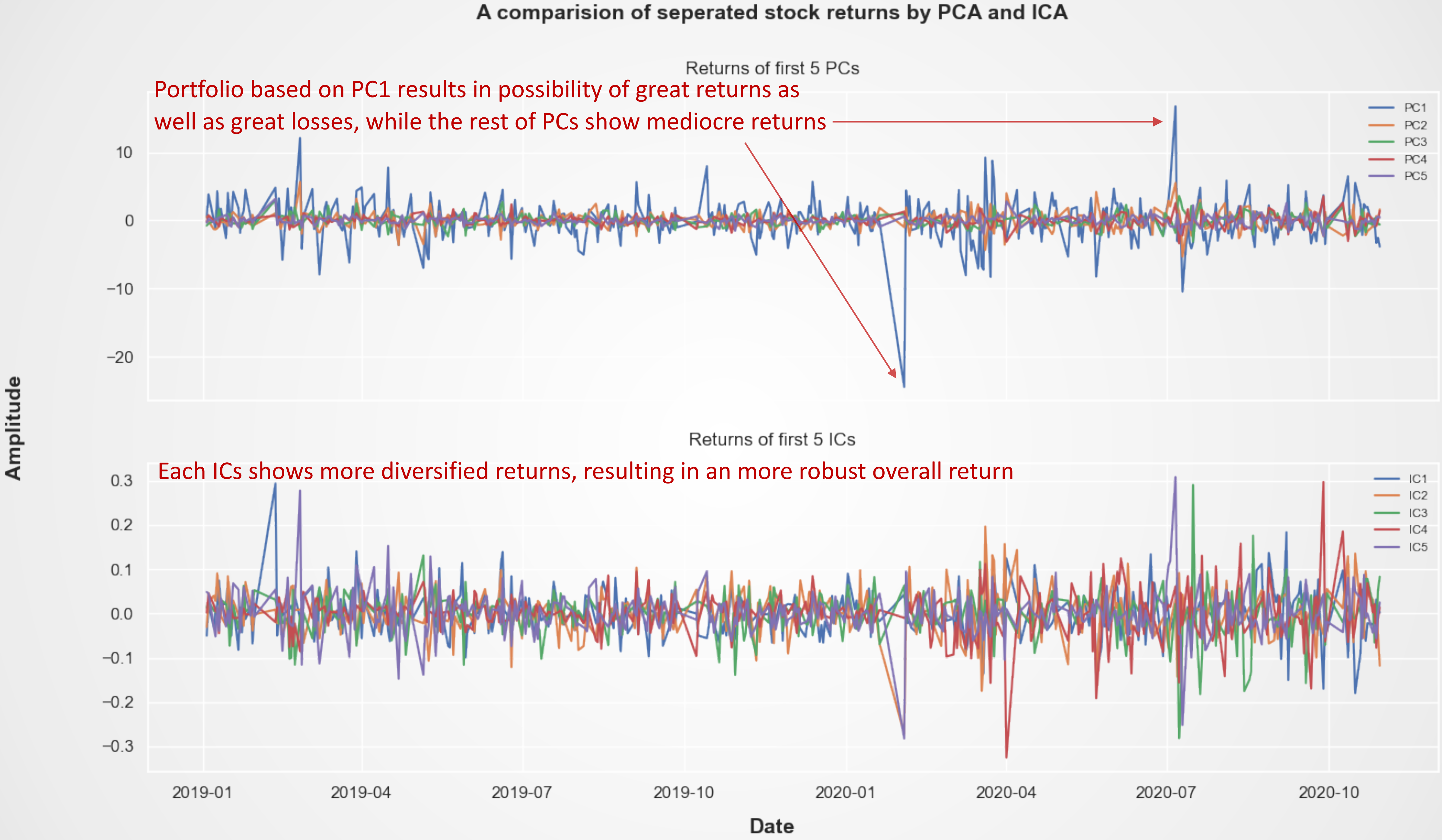
	X1398.HK	X601398.CH	X601939.CH	X939.HK	X601288.CH	X5.HK	X2318.HK	X601318.CH	X3988.HK	X601988.CH	X1299.HK	X600036.CH	X2628.HK	X601628.CH	X3328.HK
PC1	0.27	0.26	0.27	0.27	0.26	0.19	0.28	0.27	0.26	0.26	0.22	0.28	0.28	0.24	0.24
PC2	-0.25	0.27	0.28	-0.23	0.32	-0.38	-0.052	0.2	-0.25	0.27	-0.32	0.13	-0.047	0.27	-0.34
PC3	-0.29	-0.24	-0.18	-0.26	-0.27	0.017	0.28	0.32	-0.17	-0.21	0.39	0.11	0.28	0.43	-0.15
PC4	-0.27	0.12	0.14	-0.25	0.21	0.76	-0.11	-0.1	-0.2	0.15	0.22	0.043	-0.084	-0.19	-0.19
PC5	-0.15	0.049	0.1	-0.063	0.042	-0.42	0.12	0.049	-0.1	0.061	0.56	0.3	-0.28	-0.49	0.16

Portfolio with first 5 ICs

	X1398.HK	X601398.CH	X601939.CH	X939.HK	X601288.CH	X5.HK	X2318.HK	X601318.CH	X3988.HK	X601988.CH	X1299.HK	X600036.CH	X2628.HK	X601628.CH	X3328.HK
IC1	-1.3	-0.18	-0.063	-0.94	-0.19	-0.8	0.69	1.1	-0.74	-0.13	2.6	1	-0.19	-0.51	-0.68
IC2	0.95	-0.31	-0.45	0.87	-0.41	-1.2	0.39	-0.31	0.92	-0.31	0.6	-0.6	1.2	-0.51	0.87
IC3	0.2	0.59	0.74	0.26	0.56	0.11	0.2	0.38	0.11	0.46	0.16	1	-1	-2.1	0.22
IC4	-0.34	-0.037	-0.03	-0.41	0.072	3.6	-0.42	-0.44	-0.37	0.032	0.28	0.0064	-0.39	0.43	-0.13
IC5	0.035	0.79	1	0.045	0.8	0.089	0.058	0.72	-0.14	0.67	-0.87	1.1	-0.65	0.16	-0.16

- ICA framework leads to more diversified portfolios (i.e., more different weights on each component)

E.g. 3. Daily Stock Returns (15 Equities in Asia Market)



A comparision of seperated stock returns by PCA and ICA

Returns of first 5 PCs

Portfolio based on PC1 results in possibility of great returns as well as great losses, while the rest of PCs show mediocre returns

Returns of first 5 ICs

Each ICs shows more diversified returns, resulting in an more robust overall return

- Compared to stock returns based on 5 PCs, those from ICs are more dispersed

≻ reduce an investor's overall risk of permanent loss.

References

[1] Cover, T. & Thomas, J.

Elements of Information Theory

John Wiley & Sons, 1991.

Single-trial variability in event-related bold signals. NeuroImage, 15: 823-835, 2002.

[2] Hyvärinen, A.

New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit

MIT Press, pp. 273-279, 1998.

[3] Hyvärinen, A. & Oja, E.

Independent component analysis: algorithms and applications

Neural Networks, 13: 411-430, 1999.

References

[4] Hyvärinen, A., Karhunen, J., & Oja, E.

CH5. Information Theory

Independent Component Analysis

John Wiley & Sons, 2001

[5] Lee, T. W., Girolami, M., & Sejnowski, T. J.

Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources.

Neural computation, 11(2): 417-441, 1999

[6] Rosenzweig, J.

Fat Tailed Factors.

arXiv preprint arXiv:2011.13637, 2020.