Blind-source separation

and other applications of PCA & ICA

Jordan Sorokin

EEG methods - March 7th, 2017

Outline

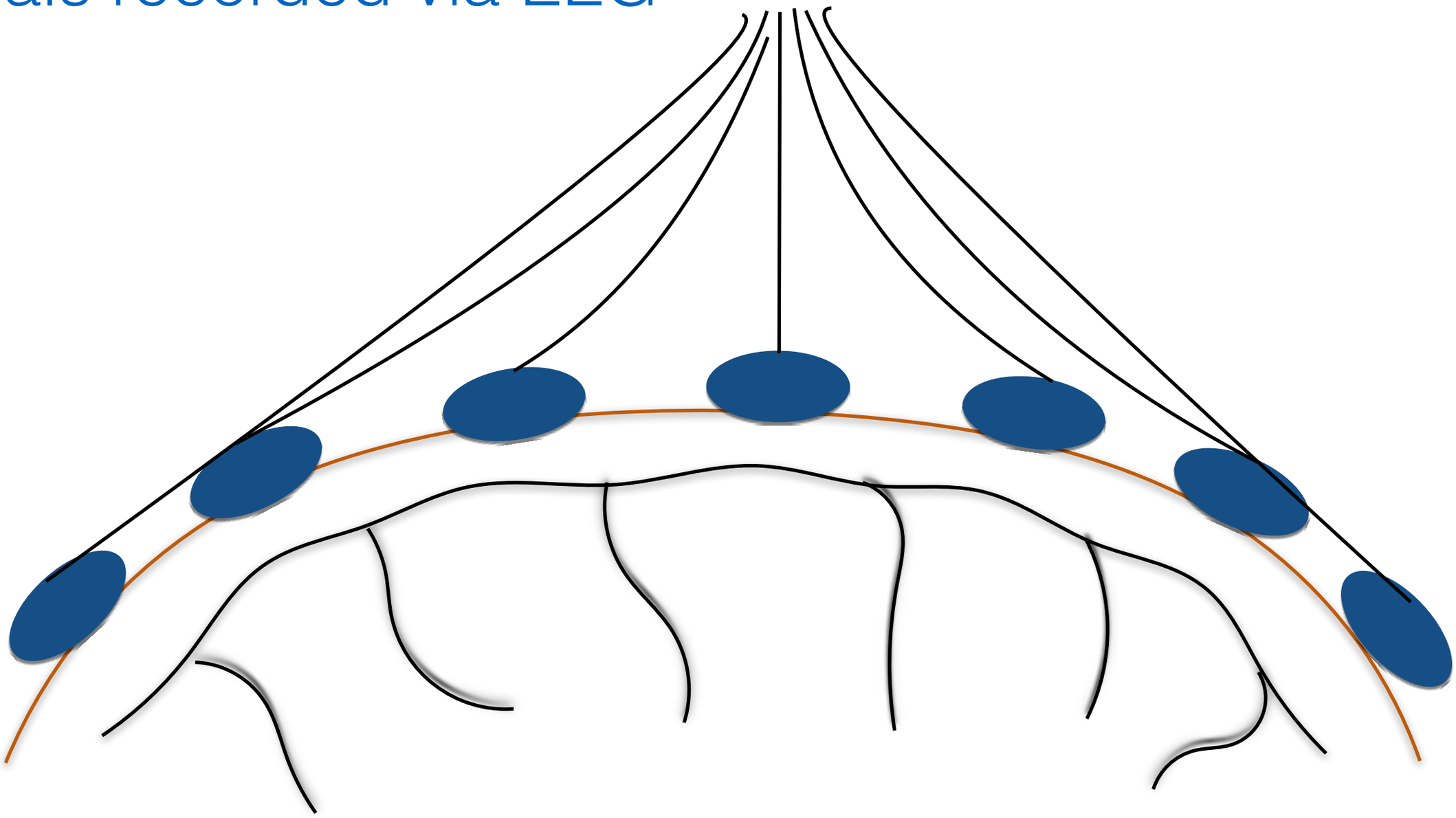
- What is source localization?
  - sources behind EEG
  - smearing & mixing of sources
  - what source separation is NOT

- Principal Component Analysis (PCA)
  - a brief review of linear algebra
  - finding a new basis (coordinate system)
  - why variance?
  - PCA in action
  - limitations

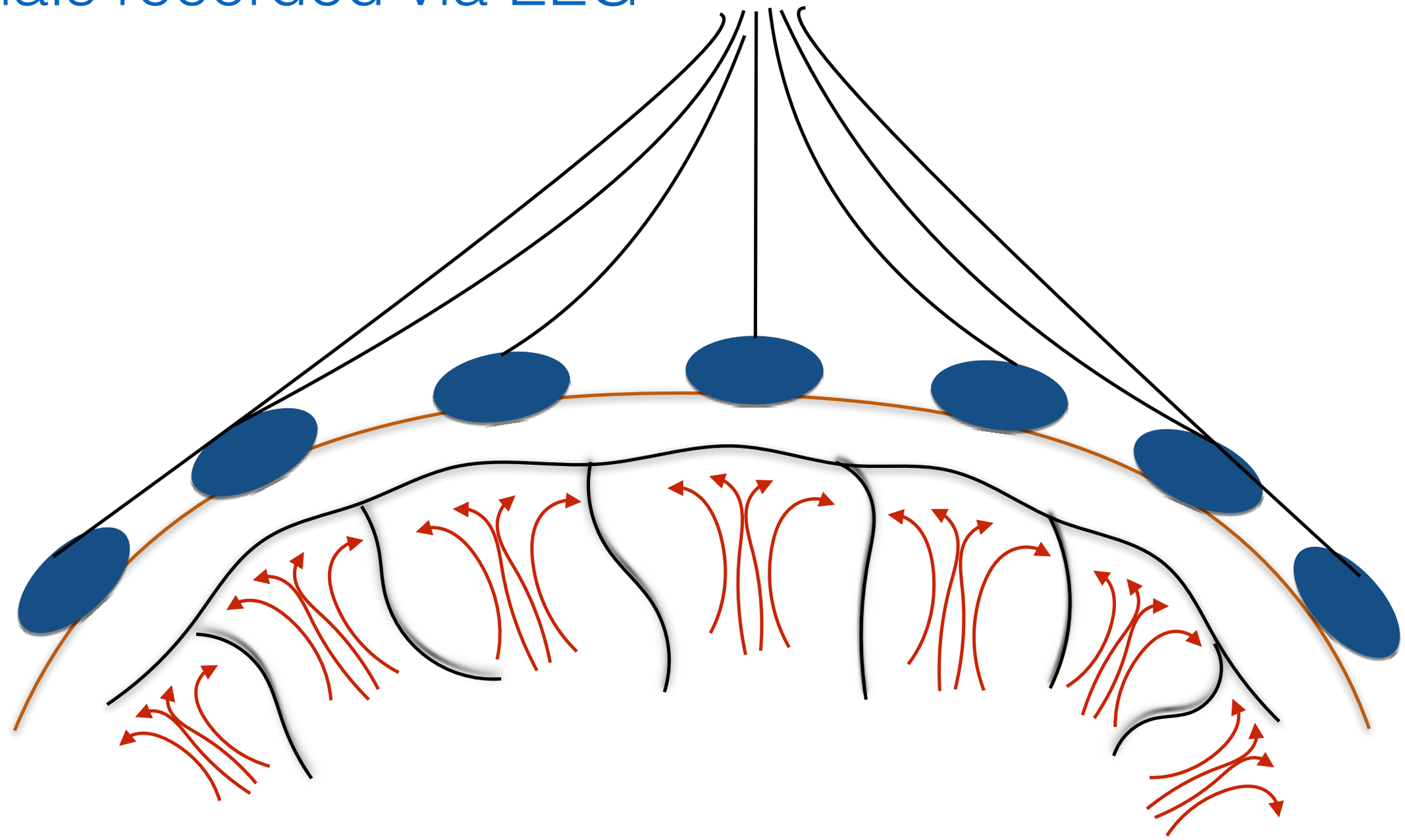
- Independent Component Analysis (ICA)
  - a revised definition of “independence”
  - method for finding a new basis
  - ICA in action
  - limitations

- Other applications of PCA/ICA
  - dimensionality reduction (projections)
  - “state-space” representation of underlying sources
  - classification / clustering
  - artifact detection / elimination

Signals recorded via EEG

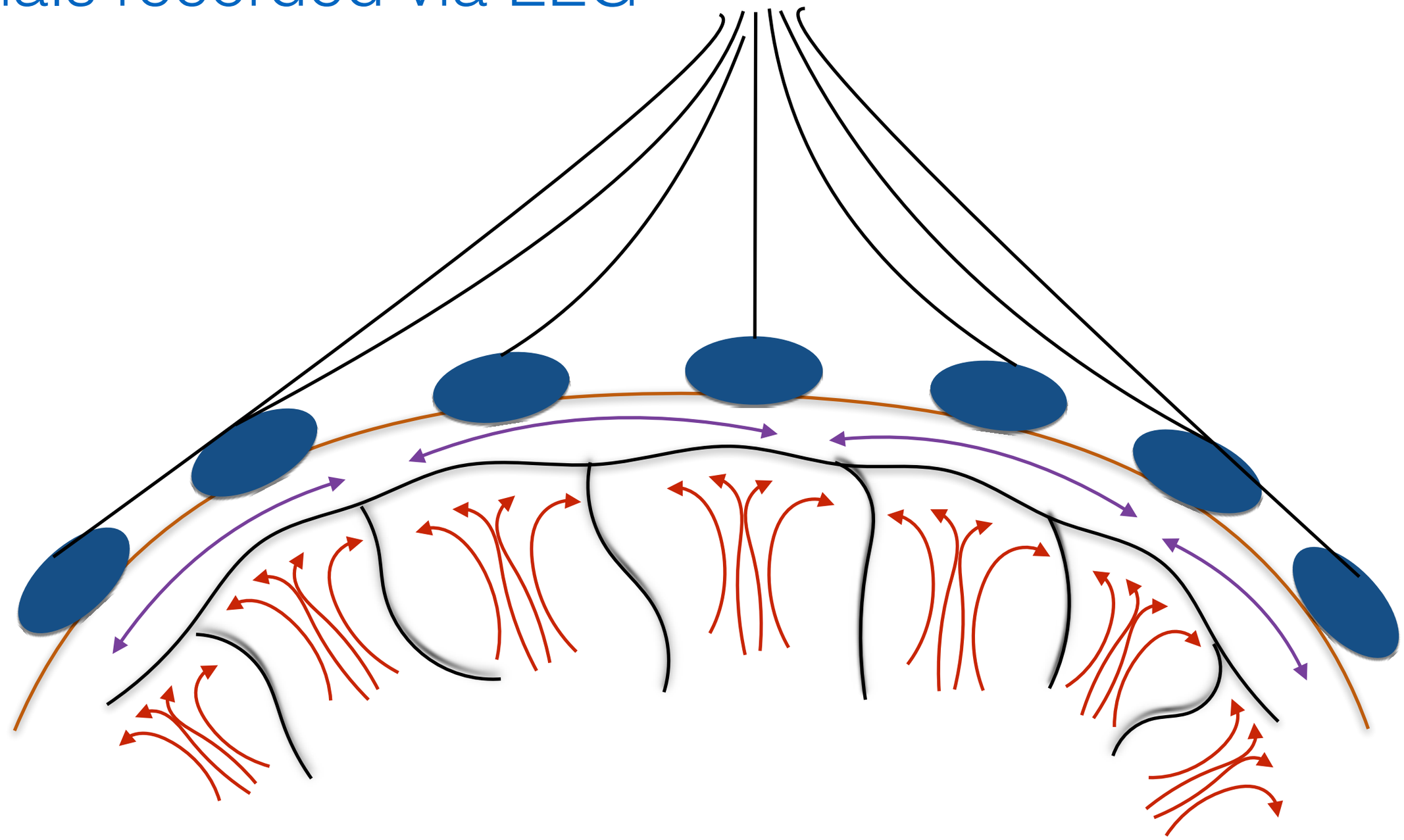


Signals recorded via EEG



(1) ascending / parallel inputs
onto upper cortical layers

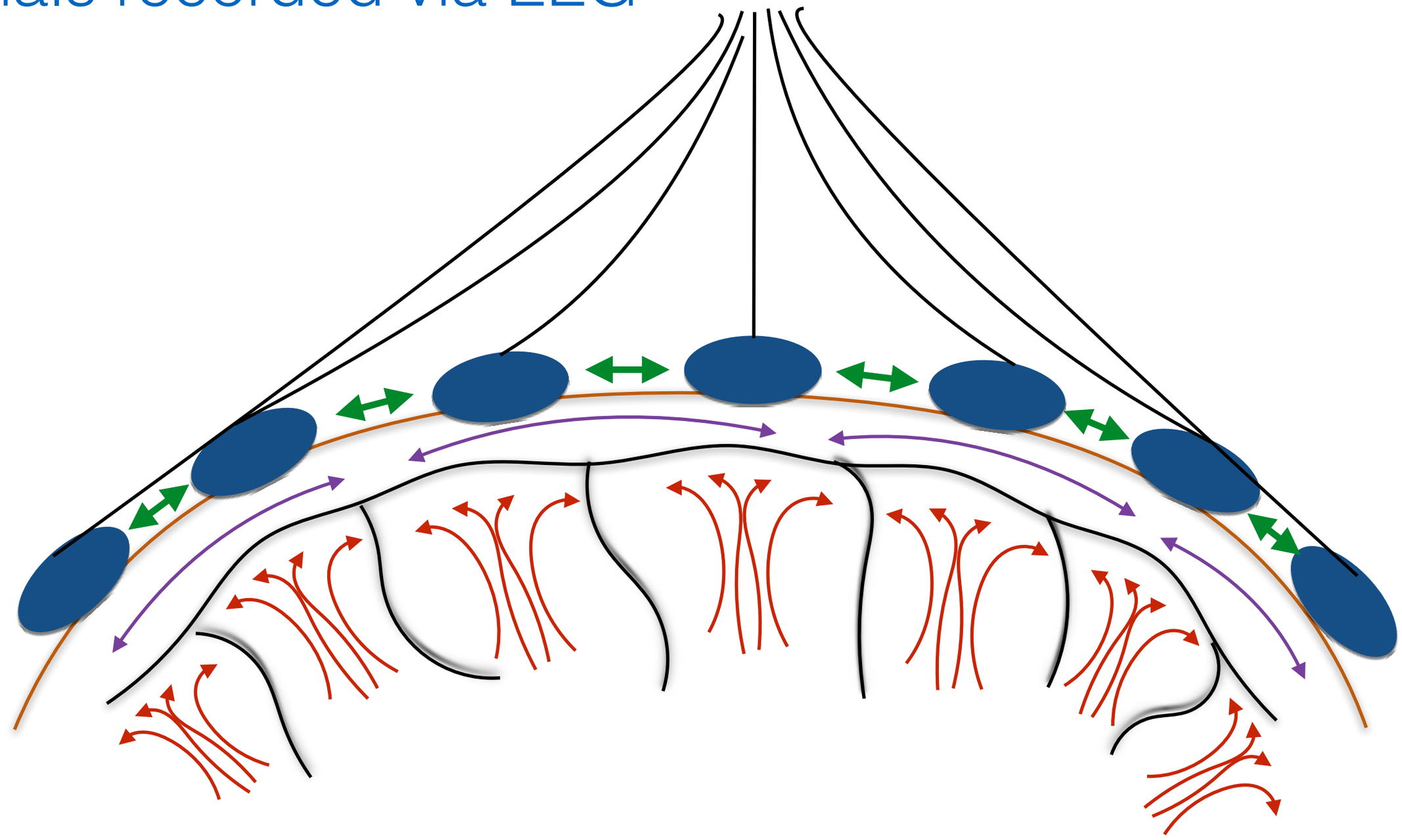
Signals recorded via EEG



(1) ascending / parallel inputs
onto upper cortical layers

(2) smearing (low pass) of signals
due to skull

Signals recorded via EEG

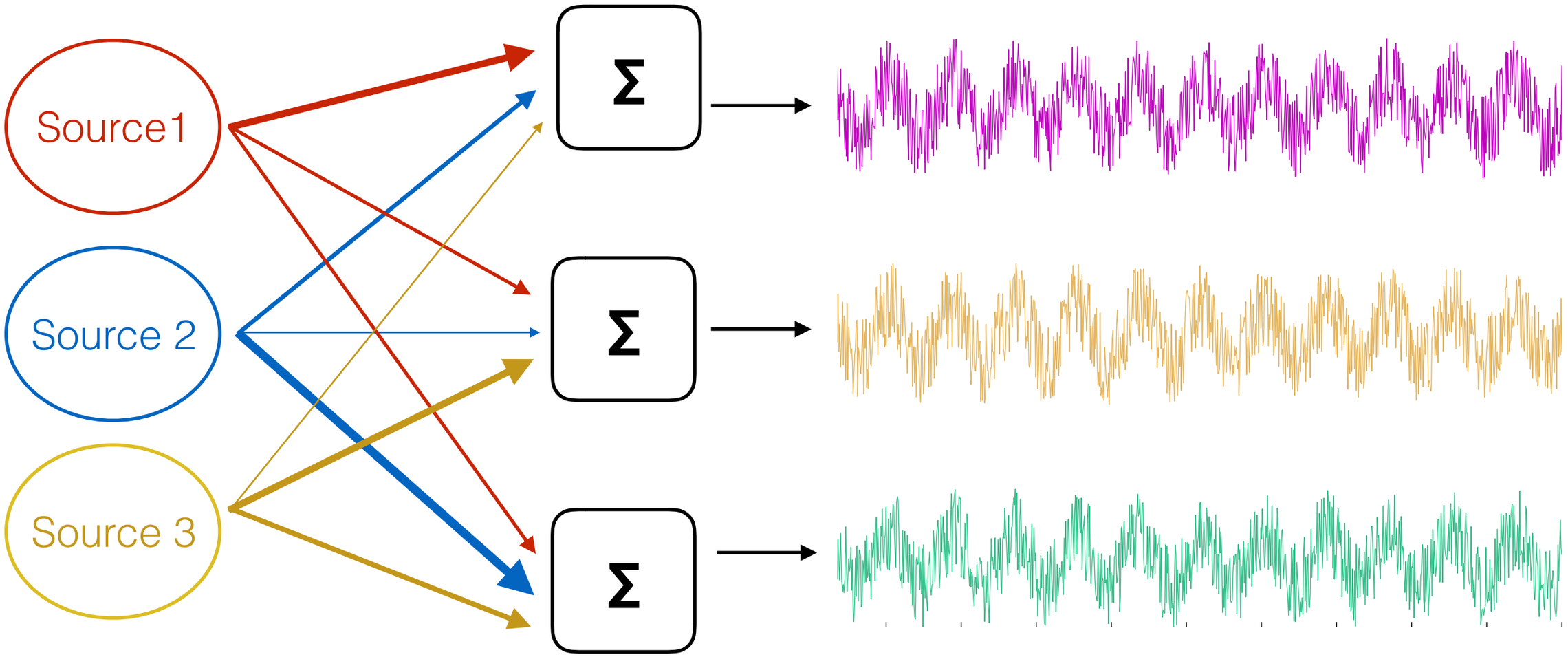


(1) ascending / parallel inputs onto upper cortical layers

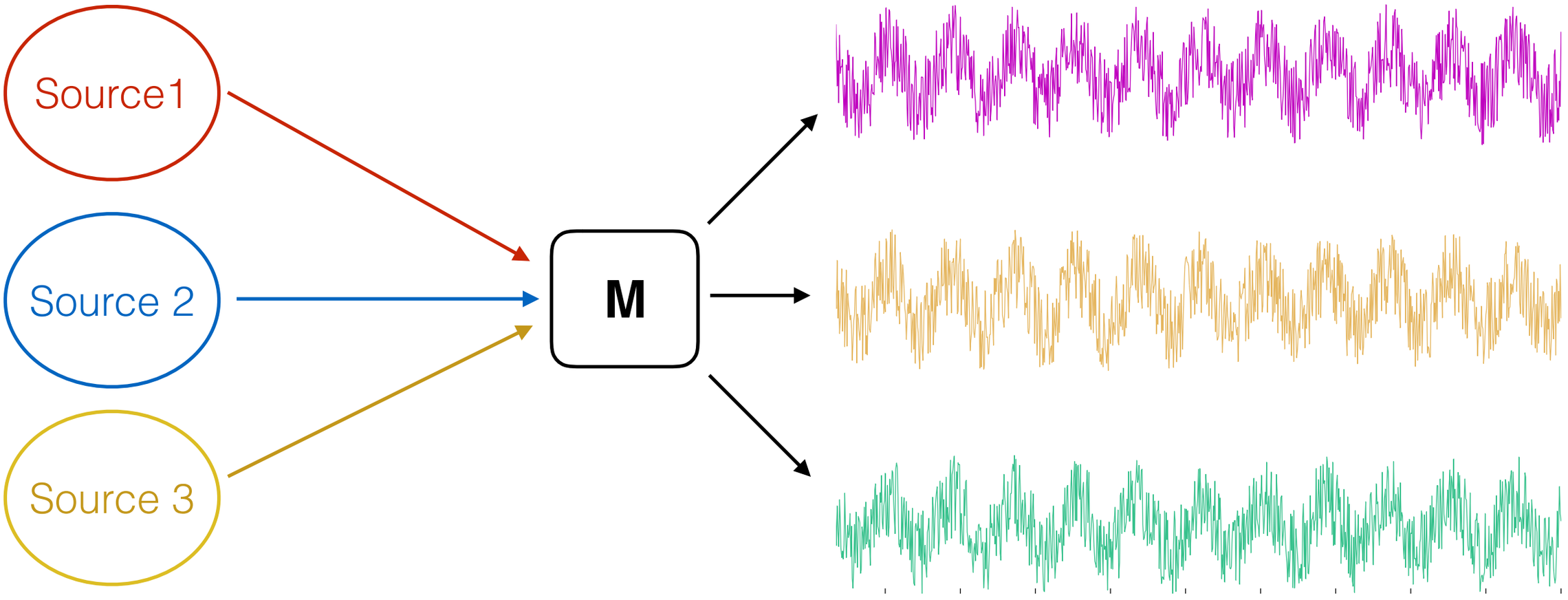
(2) smearing (low pass) of signals due to skull

(3) correlation + noise within recording electrodes

Graphical representation of source mixture

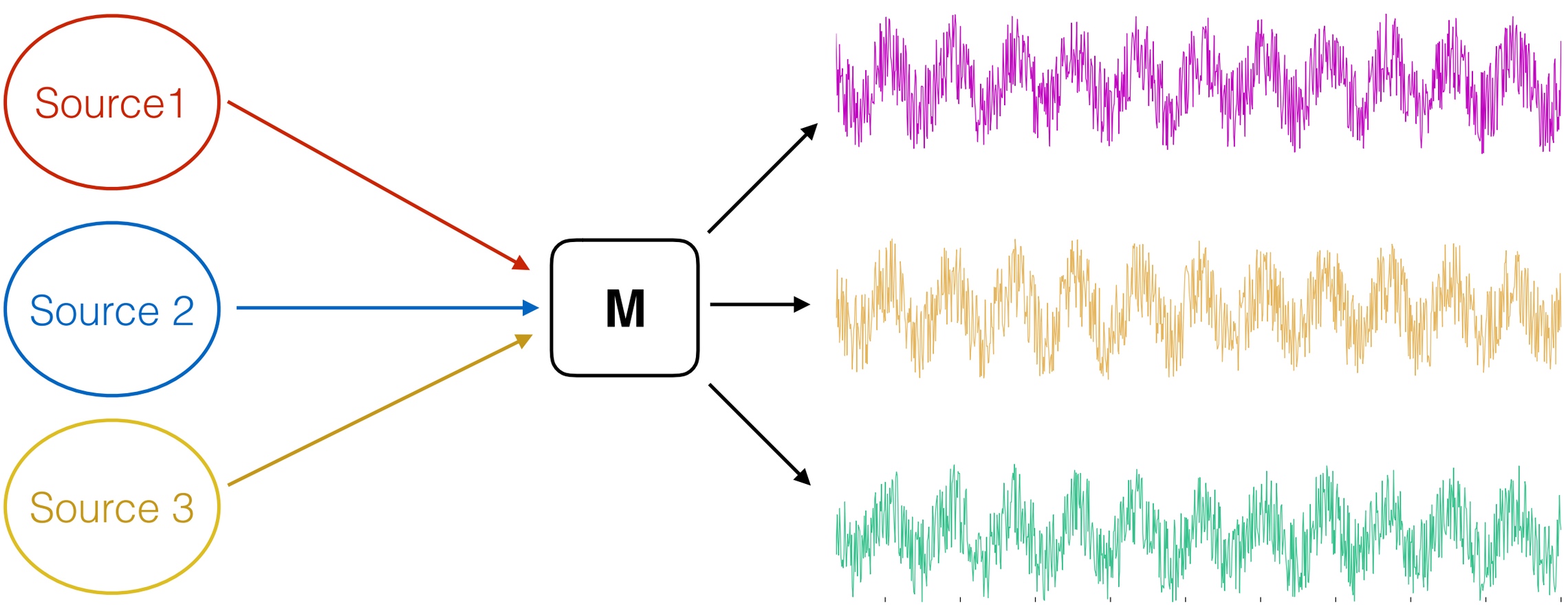


Graphical representation of source mixture



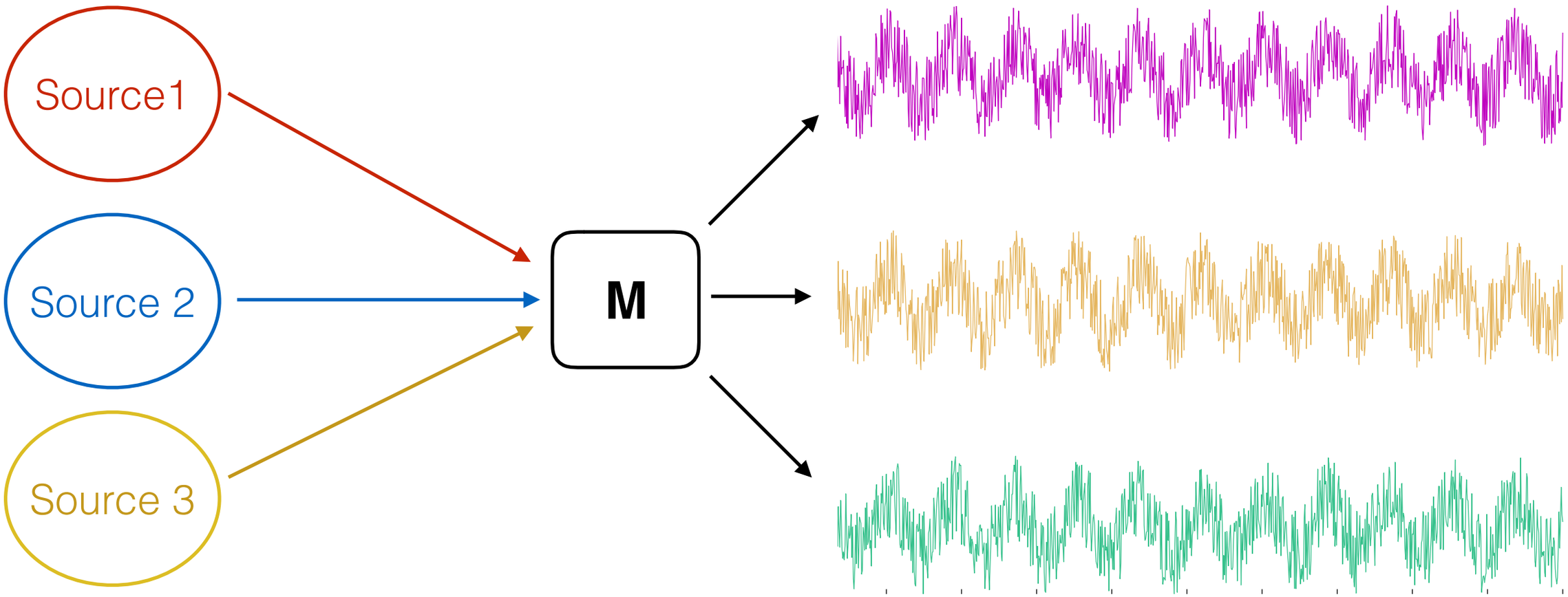
Graphical representation of source mixture

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} \\ m_{2,1} & m_{2,2} & m_{2,3} \\ m_{3,1} & m_{3,2} & m_{3,3} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \qquad X(t) = MS(t)$$



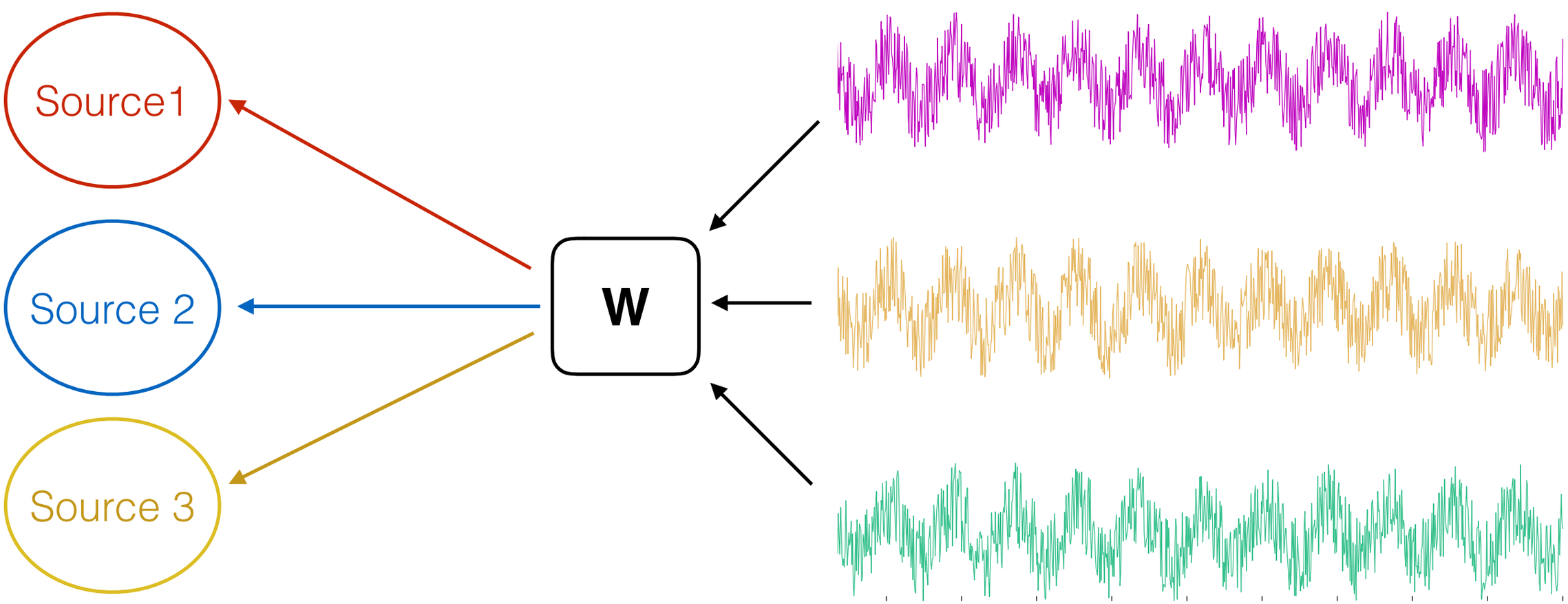
Graphical representation of source mixture

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} m_{1,1}\ m_{1,2}\ m_{1,3} \\ m_{2,1}\ m_2\ S(t) \\ m_{3,1}\ m_{3,2}\ m_{3,3} \end{bmatrix} \begin{bmatrix} s_1 \\ \\ s_3 \end{bmatrix} = WX(t)\ X(t) = MS(t)$$



Graphical representation of source mixture

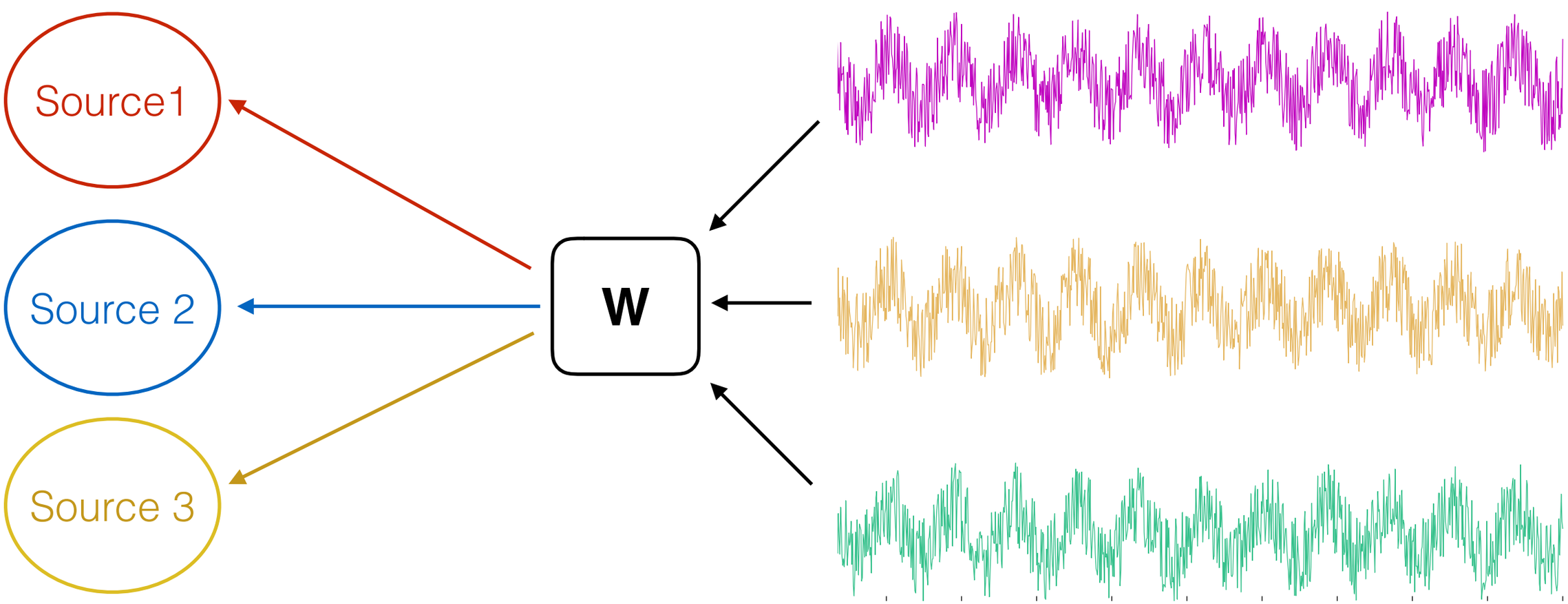
$$S(t) = WX(t)$$



What source separation is NOT

Source separation does NOT equal source localization!

Another way to put this is that source separation does not solve the inverse problem (can't locate the sources, can only estimate what they are)



Outline

- ~~What is source localization?~~
  - ~~sources behind EEG~~
  - ~~smearing & mixing of sources~~
  - ~~what source separation is NOT~~

- **Principal Component Analysis (PCA)**
  - **a brief review of linear algebra**
  - **finding a new basis (coordinate system)**
  - **why variance?**
  - **PCA in action**
  - **limitations**

- Independent Component Analysis (ICA)
  - a revised definition of "independence"
  - method for finding a new basis
  - ICA in action
  - limitations

- Other applications of PCA/ICA
  - dimensionality reduction (projections)
  - "state-space" representation of underlying sources
  - classification / clustering
  - artifact detection / elimination

Brief review of linear algebra

$$S(t) = WX(t)$$

$$s_i = [w_{i,1}\ w_{i,2}\ w_{i,3}] \bullet \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

(dot product of column of W and X)

(**Projection** of X onto W)

$$\sigma_x = \frac{1}{n}\sum (x_i x_i) \qquad \text{(variance)}$$

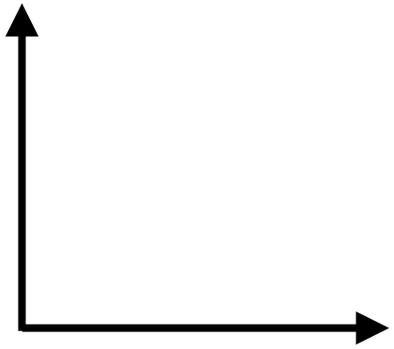
$$C_x = \frac{1}{n}\sum x_i x_j = \frac{1}{n} XX^T \qquad \text{(covariance)}$$

$$A = \begin{bmatrix} a_{1,1} & \dots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,m} \end{bmatrix} \quad A^T A = AA^T = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \qquad \text{(orthogonal)}$$

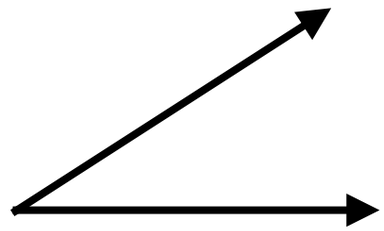
Brief review of linear algebra: orthogonality

$$A^T A = A A^T = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

orthogonal in 2D (basis)

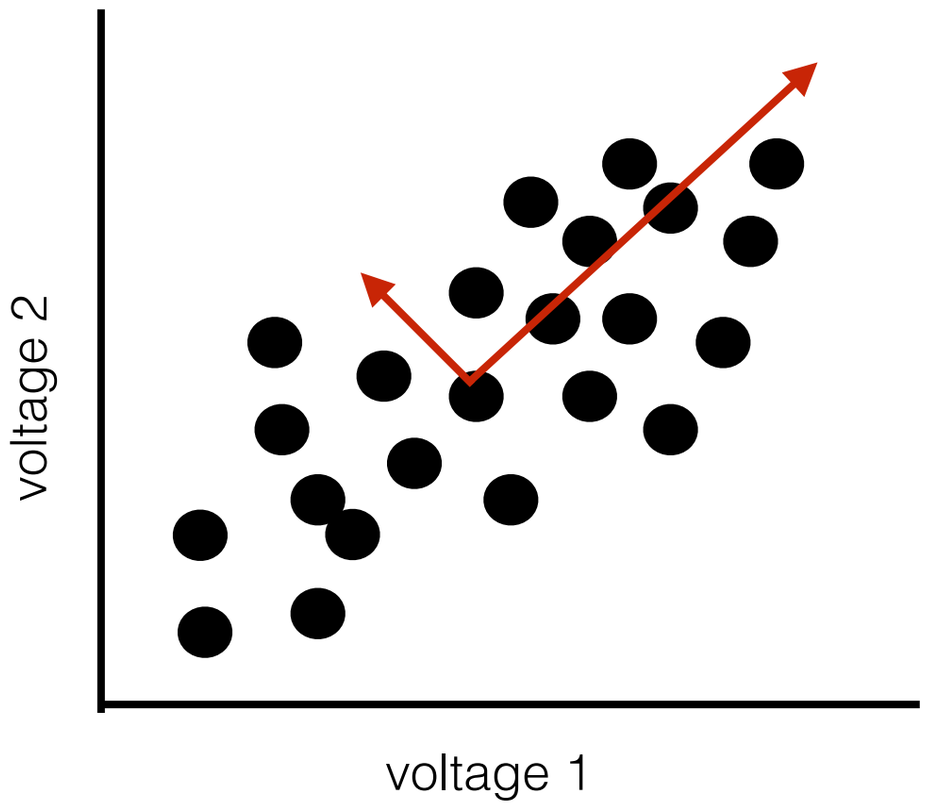


non-orthogonal in 2D



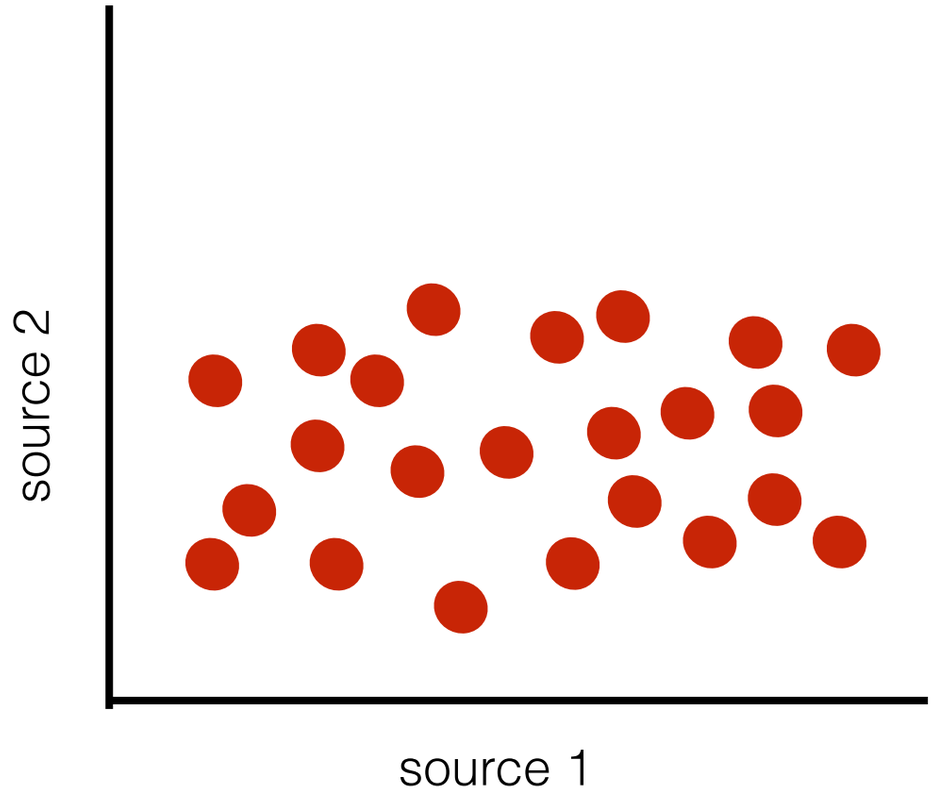
PCA tries to find a new basis for your data!

Finding a new basis (coordinate system)



we naively chose the axis of “voltage 1 vs. voltage 2” as our basis (coordinate system)

is there a better basis to project our data onto?



PCA finds a new basis that maximizes variance, minimizes covariance

Why variance?

Main motivating questions:

(1) How can we *best express* our original data X to eliminate redundancy (estimate true underlying sources)?

(2) What is the *best matrix W* to use to answer question (1) ?

The use of variance in PCA comes from answering those questions:

(1) Best express X by maximizes the variance of X (assume large variance represents "true" signals, small variance represents noise)

(2) Choose *W* to maximize variance but *minimize covariance* to remove redundancy (estimate orthogonal *or independent* sources from mixed ones)

$$C_x = \frac{1}{n} XX^T$$ ← Off-diagonals may not = 0 (redundant)

$$C_s = \frac{1}{n} SS^T$$ ← Off-diagonals = 0 (no redundancy / orthogonal)

Mathematical justification of using variance:

$$C_x = \frac{1}{n} XX^T$$

$$C_s = \frac{1}{n} SS^T$$

$$S = WX$$

substitute XX' $C_s = \frac{1}{n} WX(WX)^T$

how do we make WCxW = diagonal matrix?

(XY)' = Y'X' $= \frac{1}{n} WX(X^T W^T)$

commutative rule $= W\frac{1}{n} XX^T W^T$

definition of cov(X) $= WC_x W^T$

ECxE' = D $= W(E^T D E) W^T$

W = E $= (WW^T) D (WW^T)$

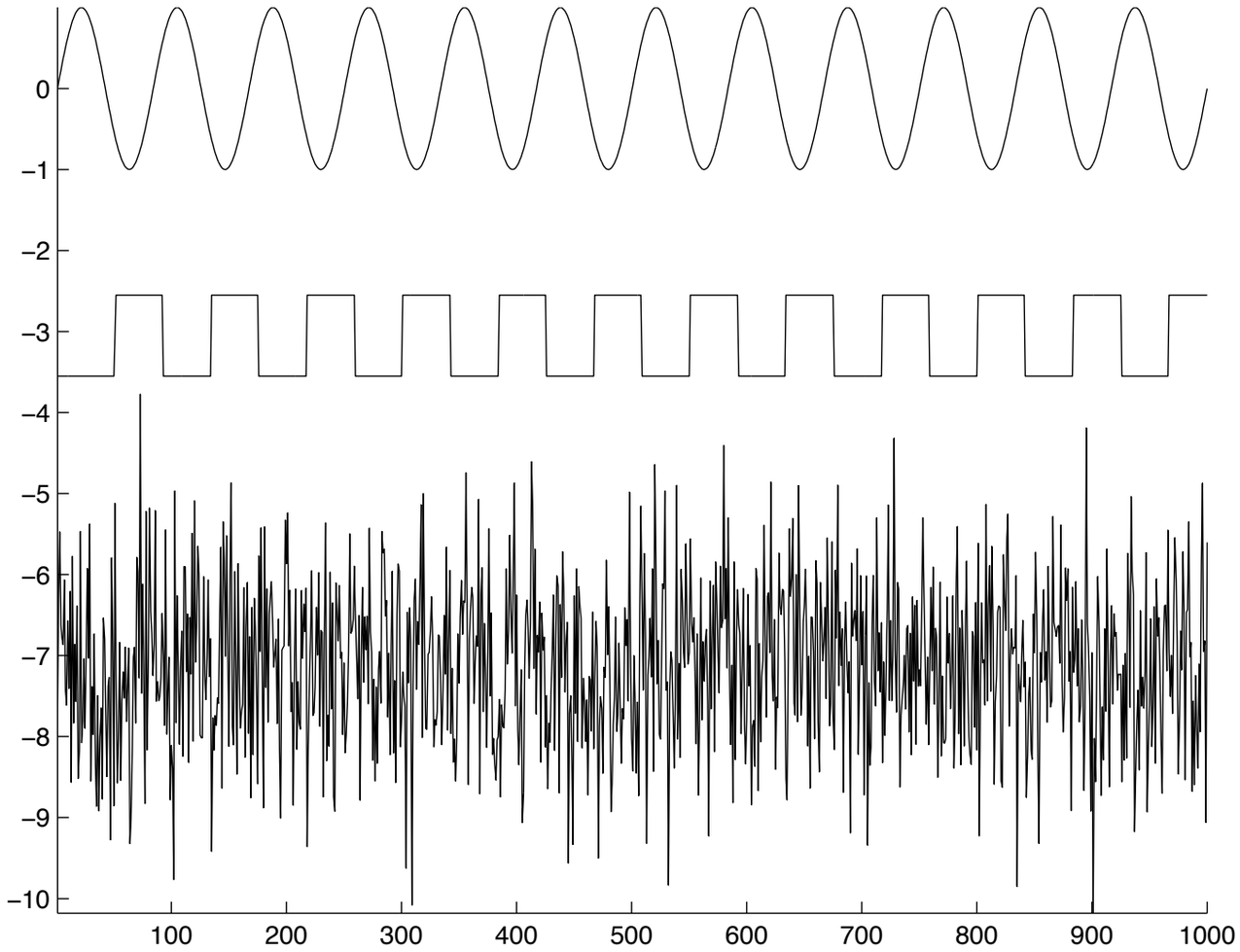
W' = inv(W) $= (WW^{-1}) D (WW^{-1})$

diagonal matrix! $= D$

We eliminate covariances in our data by finding the eigenvectors / values of Cx!

PCA in action:

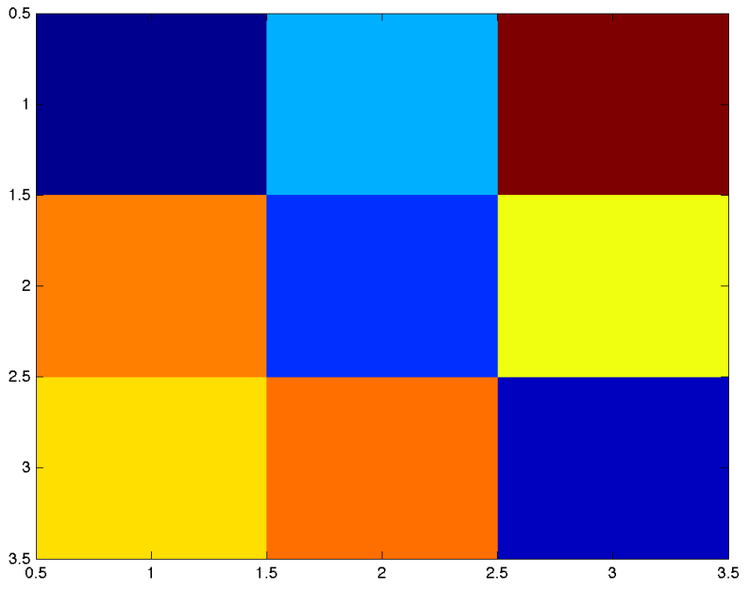
```
%% create the underlying sources
t = linspace( 0,1,1000 ); % time vector
S = zeros( 3,1000 ); % the source matrix
S(1,:) = sin( 2*pi*12*t ); % 12 Hz sine wave
S(2,:) = heaviside( S(1,:) ); % square wave
S(2,:) = [zeros( 1,50 ),S(2,1:950)]; % offset
S(3,:) = randn( 1,1000 ); % gaussian white-noise
```

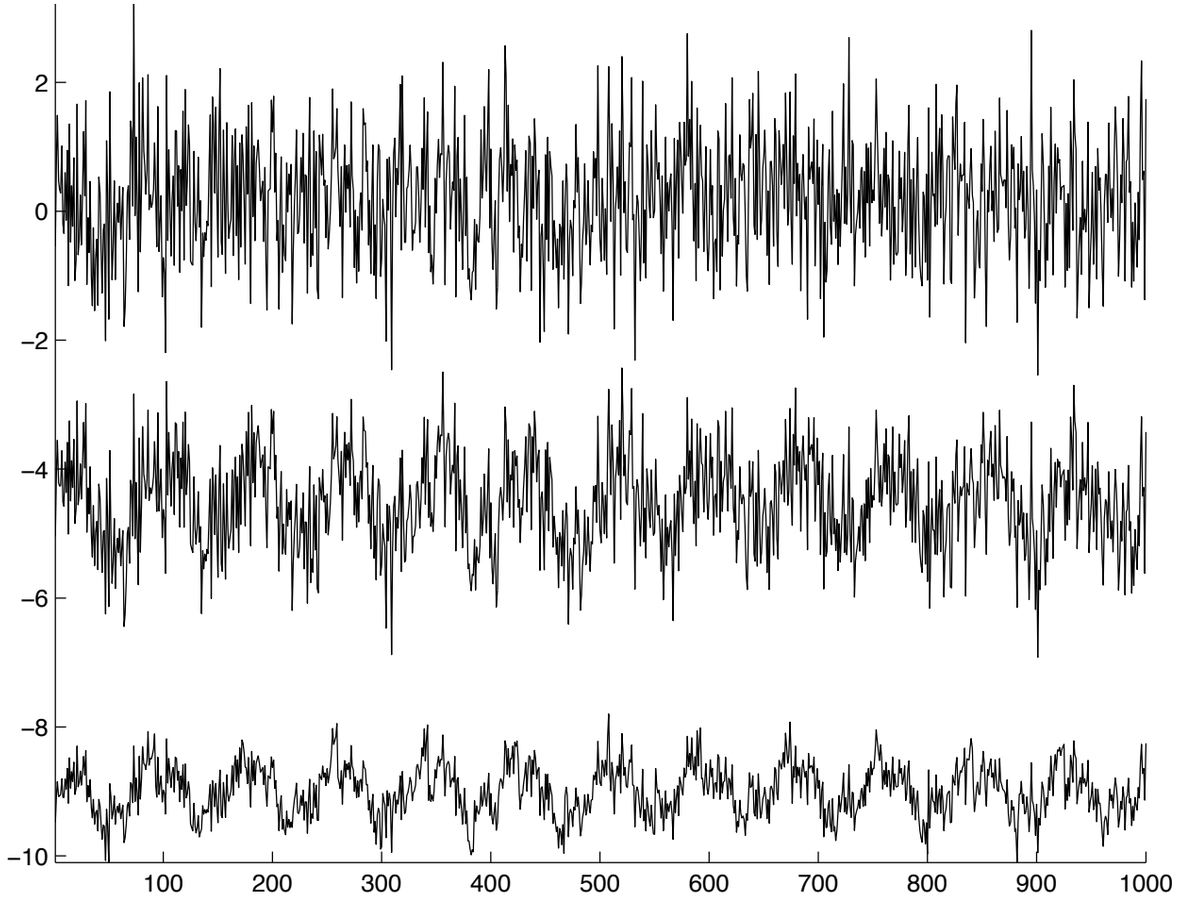


PCA in action:

```
%% create mixing matrices and observed signals
M = [0.2 0.4 0.9;...
     0.5 0.22 0.43;...
     0.67 0.75 0.25];
% normalize each row in M by it's norm, so that the final mixing of signals
% produces outputs with equal energy
for i = 1:3
    M(i,:) = M(i,:) / norm( M(i,:) );
end

X = M*S;
```

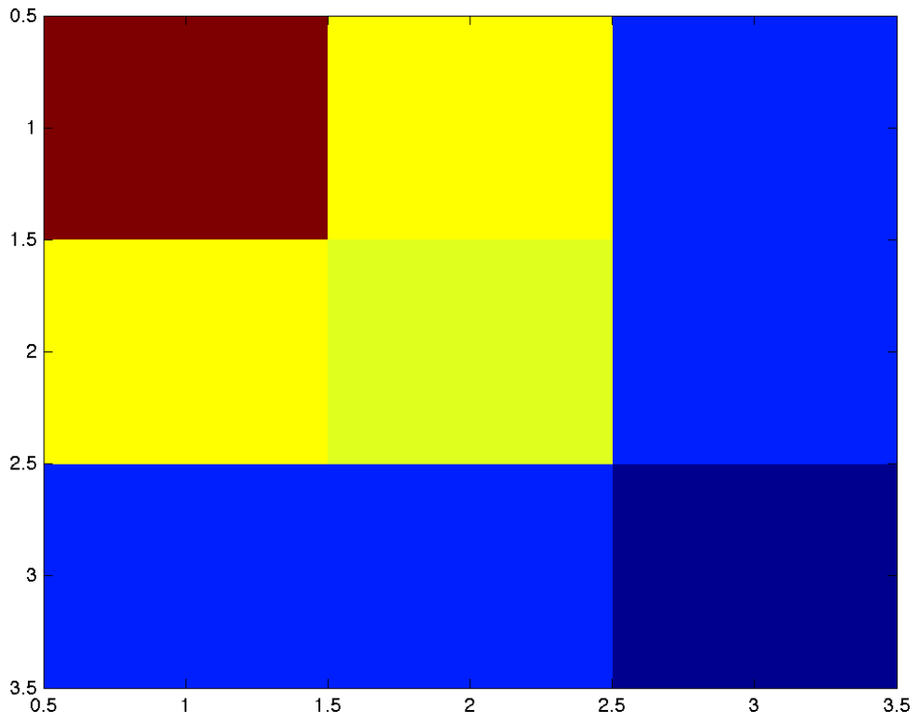


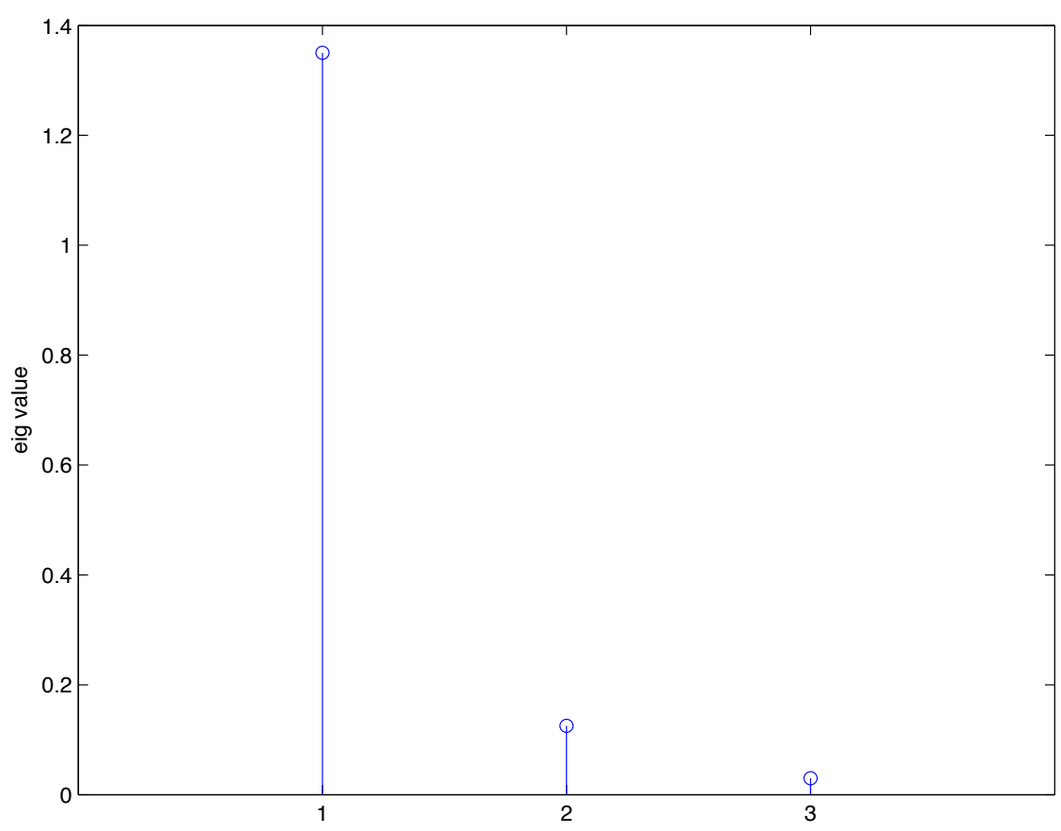


PCA in action:

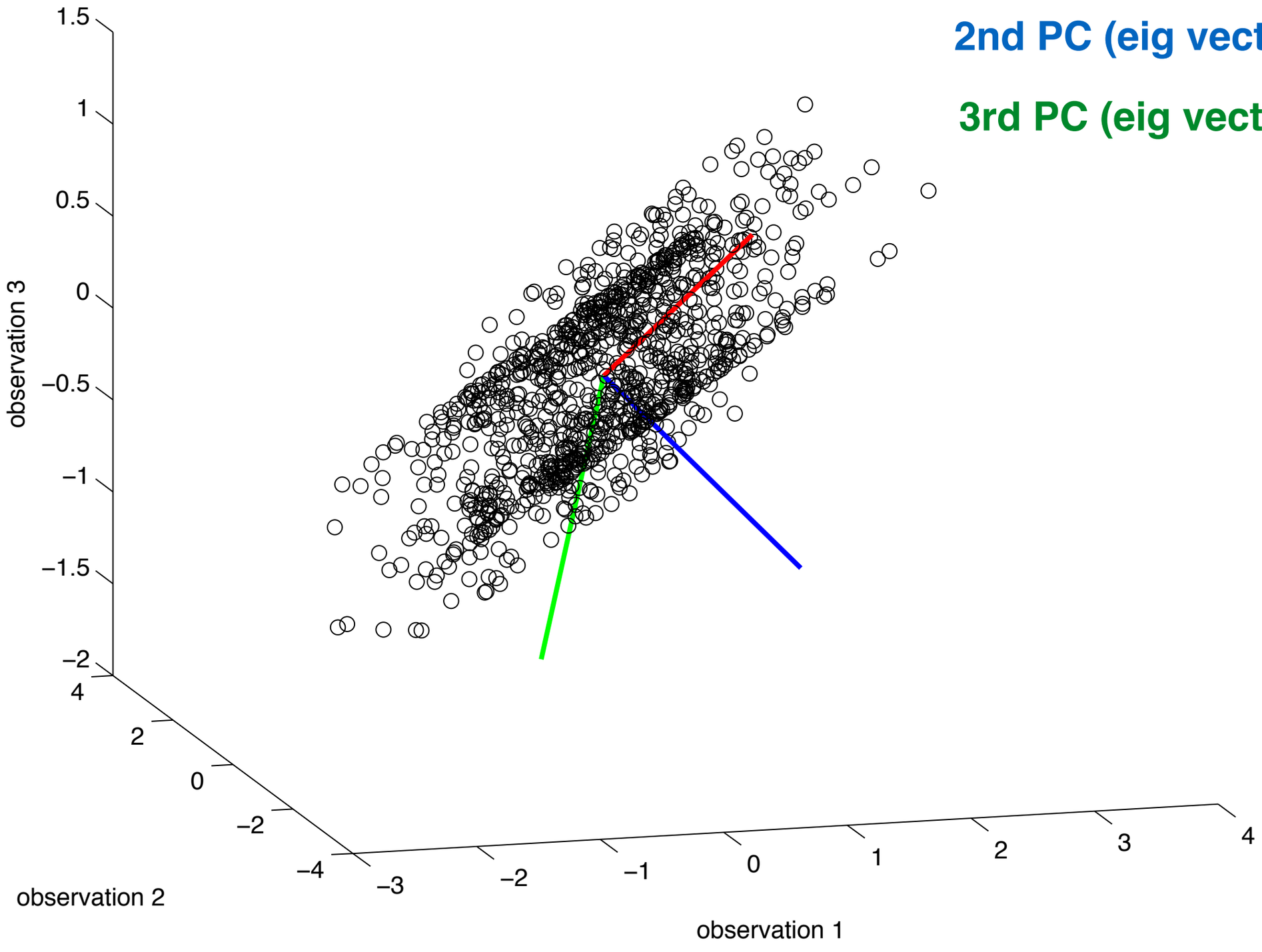
```
%% perform PCA on the mixed signals X
X = bsxfun( @minus,X,mean(X,2) ); %de-mean the time series
C = cov( X' ); % note, cov() takes in column-major time-series
               % equivalent to: 1/n * X*X' (looking at covariances across
               % channels, NOT time points)
```

```
[PC,eigval] = eig( C ); % get the principal components (eigen vectors)
eigval = diag( eigval );
[~,id] = sort( eigval,'descend' ); % sort from largest --> smallest
eigval = eigval(id);
PC = PC(:,id); % sort the PC's to correspond with the sorted eig values
```



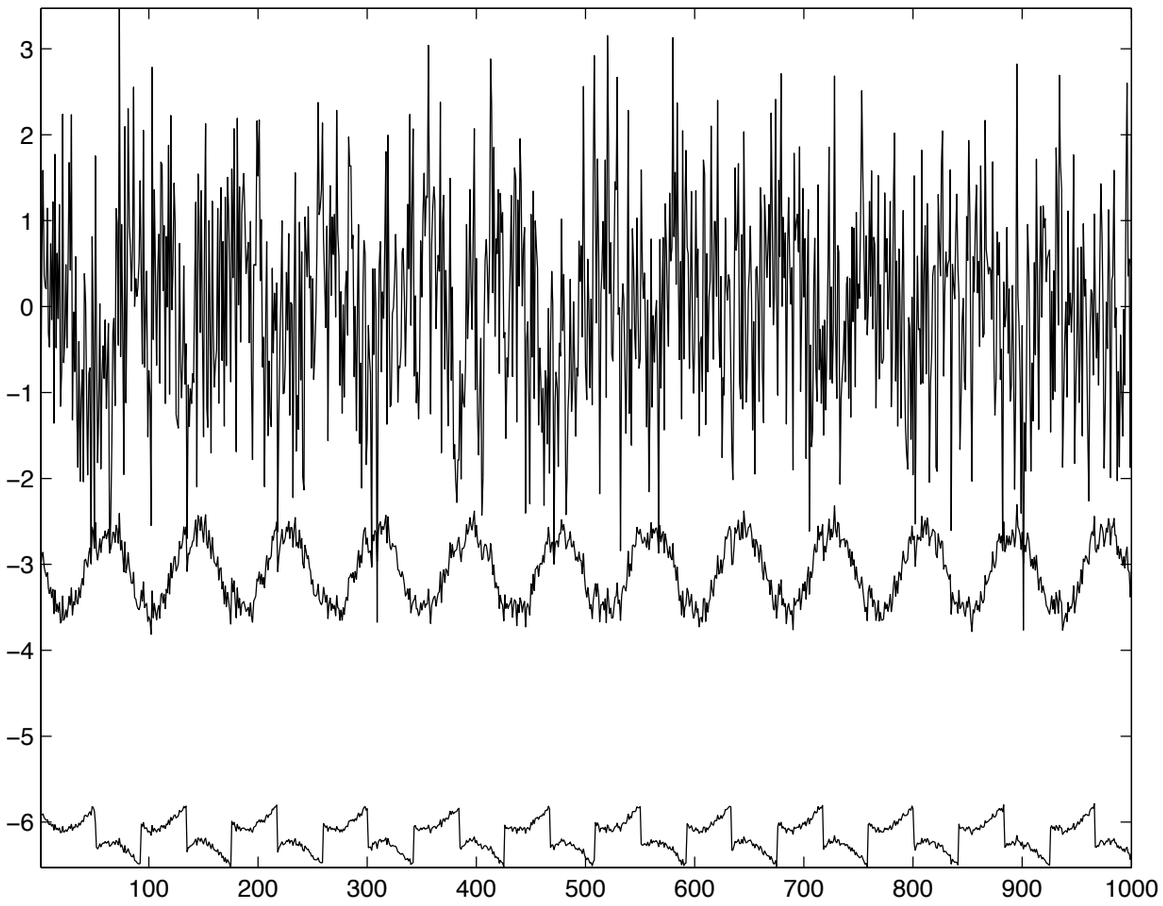


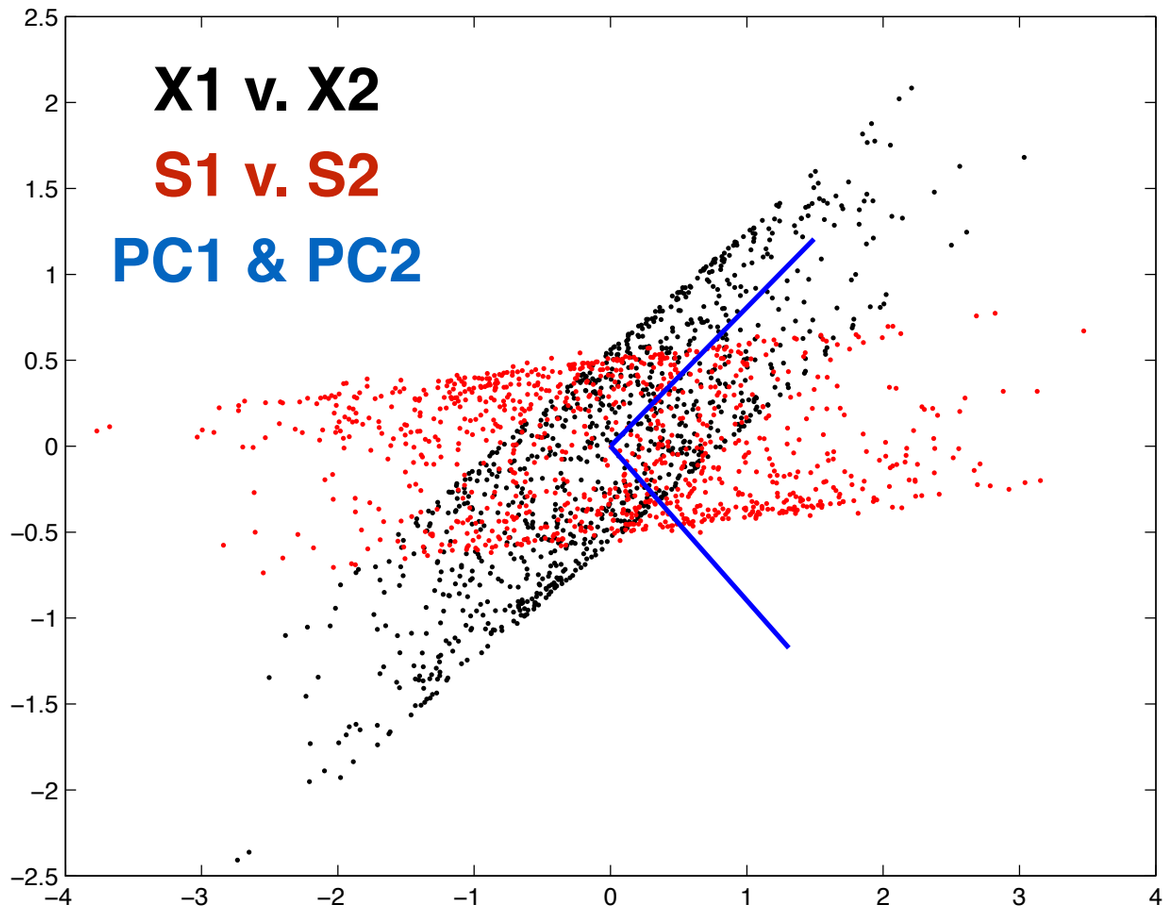
PCA in action:



PCA in action:

```
% project onto the eigenvectors/PCs (i.e. dot product with columns)
sources = PC' * X; % S = WX
```



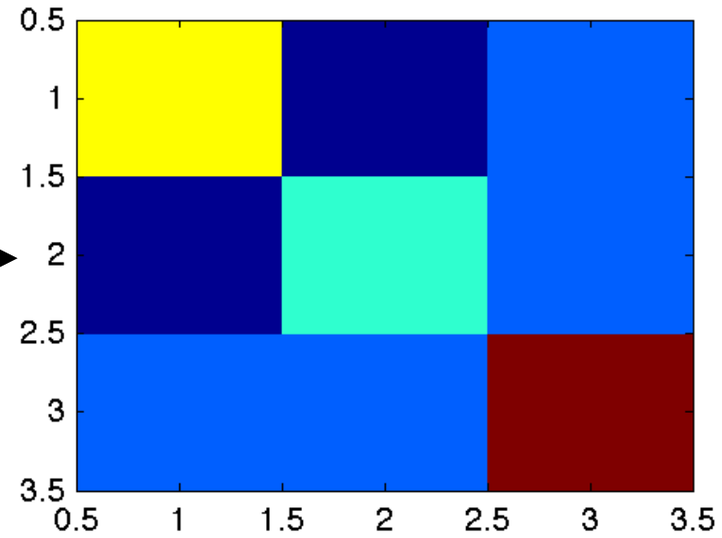


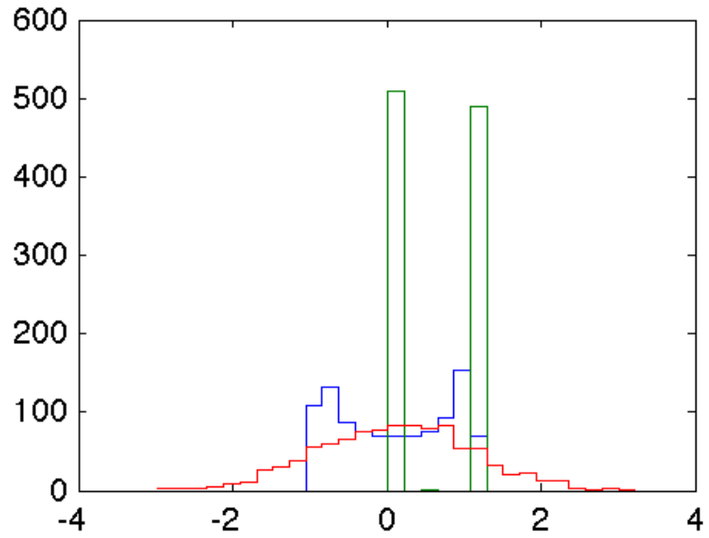
Doesn’t fully recover the sources…why?

Limitations of PCA

PCA assumes:
(1) sources are orthogonal, and
(2) data/sources are gaussian (i.e. variance/covariance is sufficient to explain the distribution of the data)

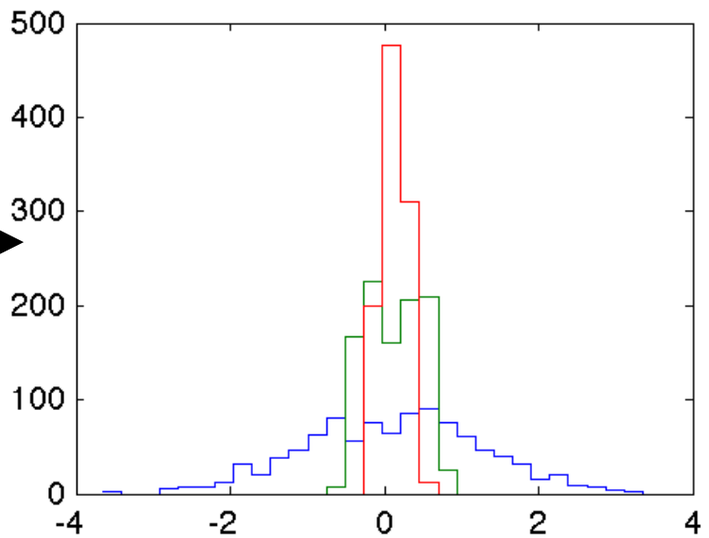
true sources
not orthogonal

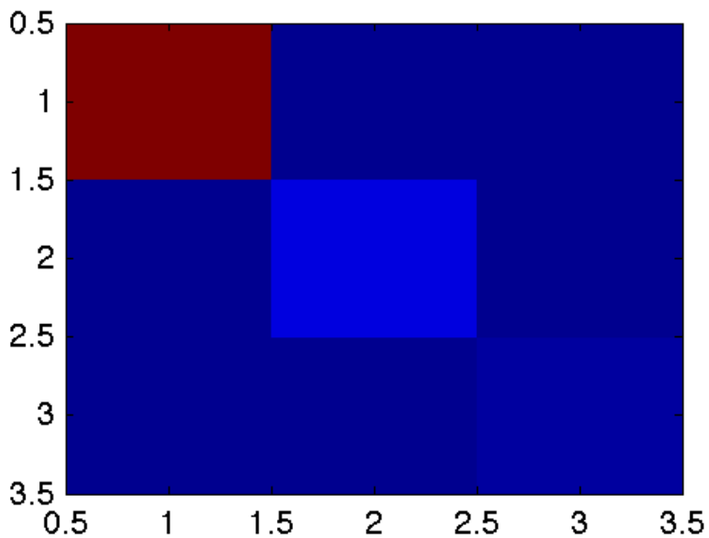




true sources
not all gaussian

estimated
sources forced
gaussians



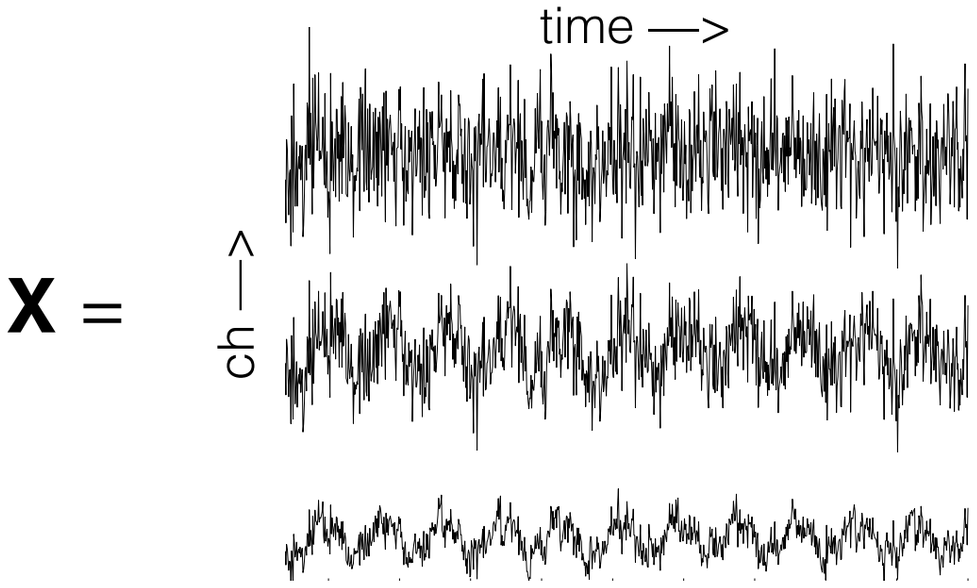


estimated
sources forced
orthogonal

Interpreting PCA

Our interpretation of PCA is highly dependent on the rows vs. columns of our input data X.

Here, we had rows = channels, columns = time. Thus, our covariance XX' represents covariance across <u>channels</u>, and our <u>PCs are of dim: # ch</u>



1/n XX' = channel covariances

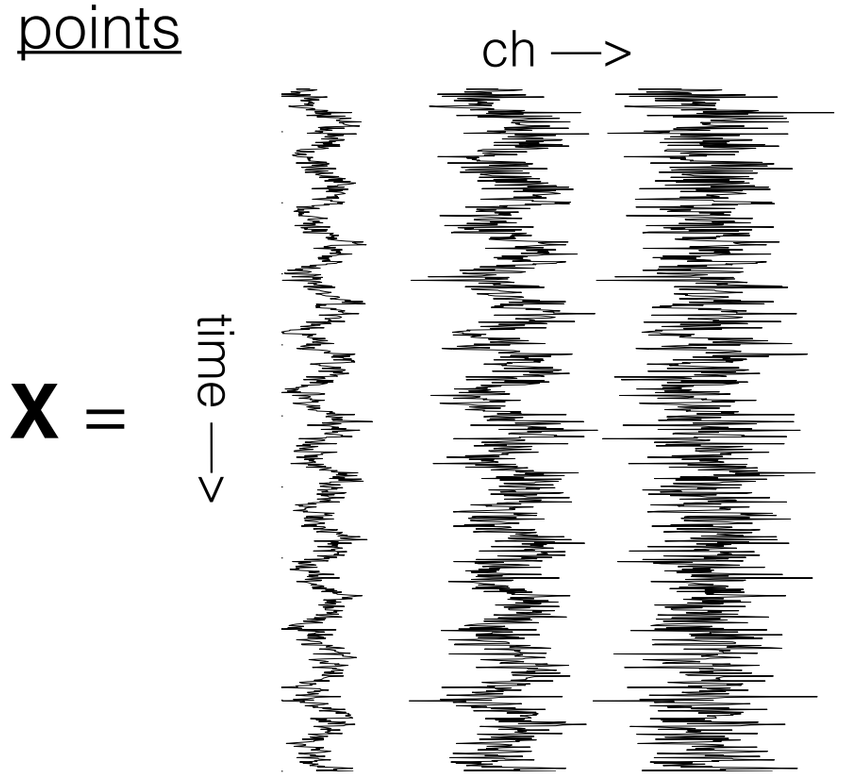
Eigvectors (PCs) = directions (weights) that describe the variance across our channels

S = PC'X = demixed time-varying sources describing the underlying variance in X

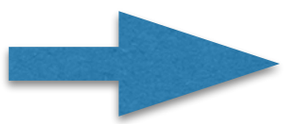
Interpreting PCA

Our interpretation of PCA is highly dependent on the rows vs. columns of our input data X.

However, if we transposed X first, then the covariance is across <u>time-points</u> (observations) and the <u>PC's represent time-varying sources of dim: # time points</u>



X =



1/n XX' = time-point covariances

Eigvectors (PCs) = time-varying features found in X

S = PC'X = projections of X onto each time-varying feature (i.e. "score" of each X onto each PC, or the amount of relatedness between each feature and X)

Outline

- ~~What is source localization?~~
  - ~~sources behind EEG~~
  - ~~smearing & mixing of sources~~
  - ~~what source separation is NOT~~

- ~~Principal Component Analysis (PCA)~~
  - ~~a brief review of linear algebra~~
  - ~~finding a new basis (coordinate system)~~
  - ~~why variance?~~
  - ~~PCA in action~~
  - ~~limitations~~

- **Independent Component Analysis (ICA)**
  - **a revised definition of "independence"**
  - **method for finding a new basis**
  - **ICA in action**
  - **limitations**

- Other applications of PCA/ICA
  - dimensionality reduction (projections)
  - "state-space" representation of underlying sources
  - classification / clustering
  - artifact detection / elimination

A revised definition of “independence”

For PCA, we observed that it forced *orthogonality* as a criteria for independence.
Another way to state this: 2nd-order (variance) decorrelation of X

In general, this is a weak interpretation of independence that may not recover true sources in our data (as we observed).

A more stringent definition of independence stems from probability theory…

$$P(x,y) = P(x)P(y) \qquad X_1\ X_2\ \ldots\ X_M \sim f_1(x)\ f_2(x)\ \ldots\ f_M(x)$$

$$Y = \sum_{j=1}^{M} X_j \sim \mathrm{N}(\mu,\sigma^2)$$

So if adding many independent non-gaussian variables results in a gaussian…

…then maximizing independence = minimizing “gaussianity” of the underlying sources!

How to minimize “gaussianity”

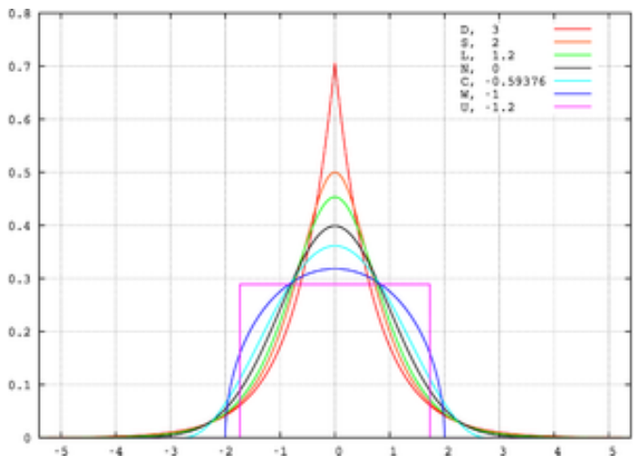
A gaussian variable can be completely described by it’s *mean* and *variance*

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

However, non-gaussian distributed variables are described by *higher-order stats*

$$\mathrm{M}^k = \mathrm{E}\left[(X - \mu_X)^k \right] = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)^k$$

M^4 *(kurtosis)*



Thus we can minimize the “gaussianity” of sources by maximizing their non-gaussian kurtosis

$$\left| K(S_i) - K(\mathrm{N}) \right|$$

How to minimize “gaussianity”

In reality, trying to maximize $|K(S_i) - K(\mathrm{N})|$ is unstable as kurtosis is sensitive

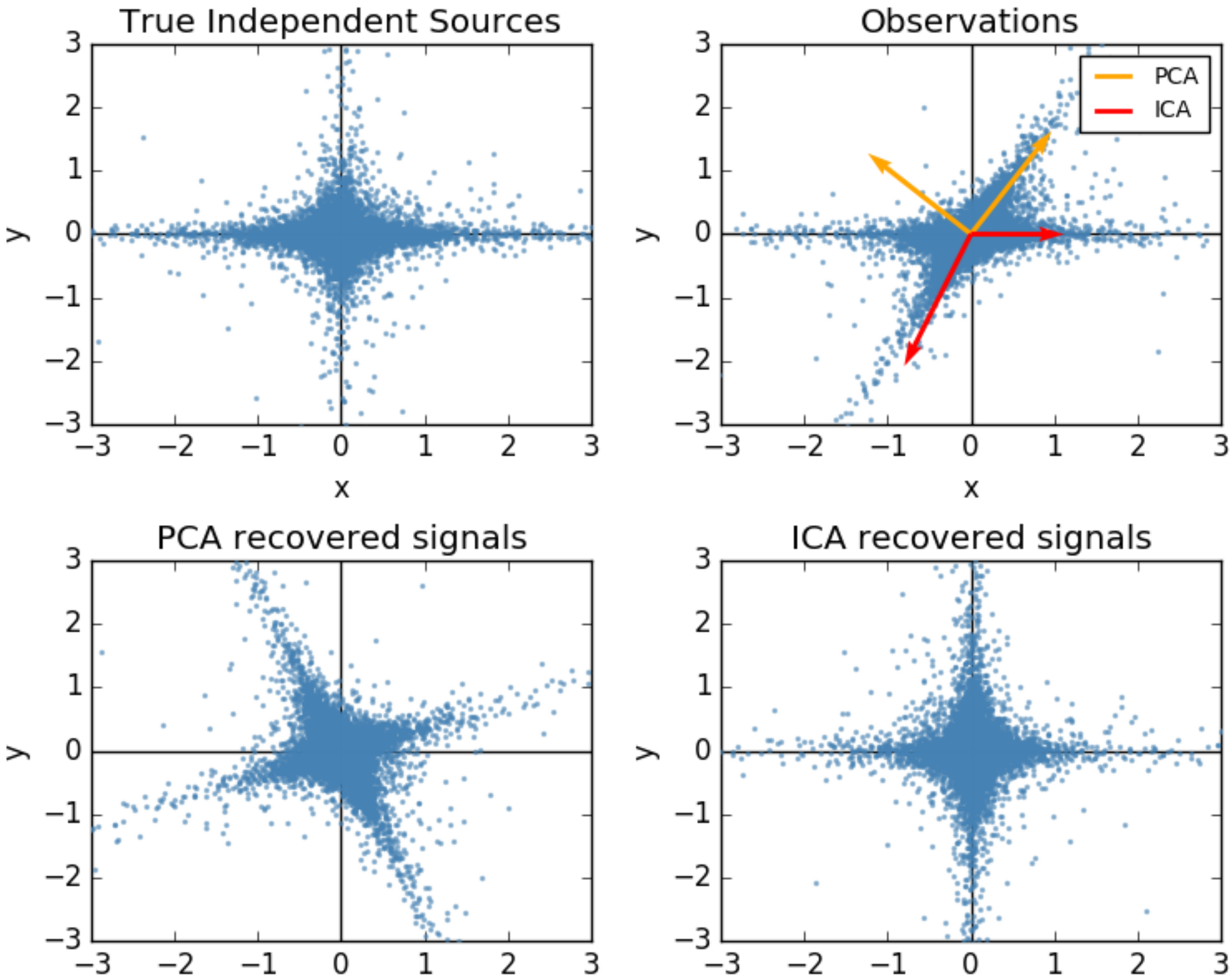
Instead, we can use other methods to try to make the sources as non-gaussian as possible, all of which are *related to kurtosis*

Biomedical Signal and Image Processing Spring 2005 , Ch. 15

Because ICA does not impose decorrelation, the new basis vectors may not be orthogonal.

…means higher-order statistics between variables are taken into consideration, and sources not imposed gaussian.

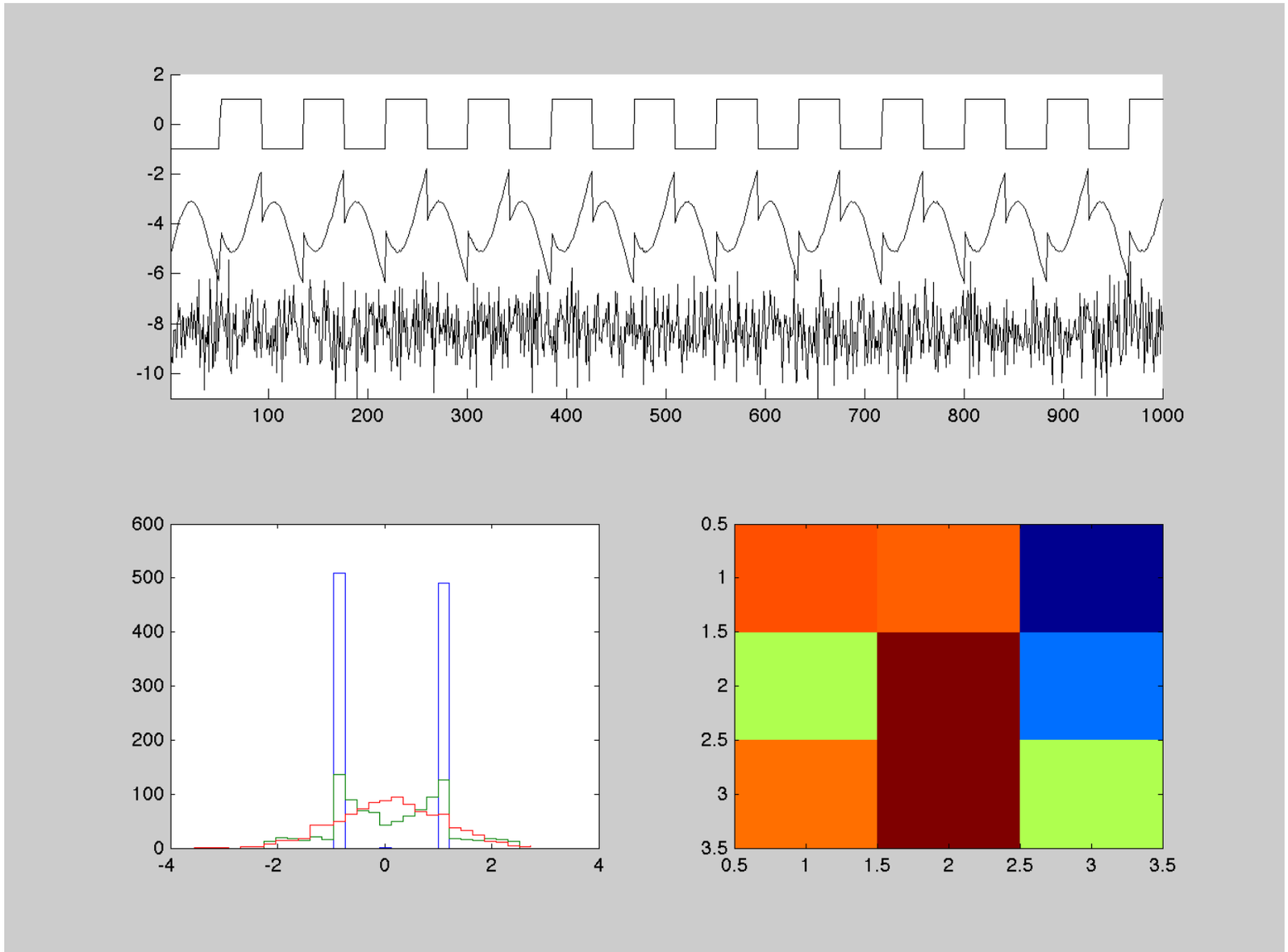
A revised definition of “independence”



ICA in action

sources are better captured, however not perfectly…

…perhaps because one of the sources is gaussian-distributed



ICA limitations

Although ICA imposes less assumptions than PCA, it still:

(1) assumes sources are independent
(2) assumes sources are as non-gaussian as possible that, when combined, create gaussian-distributed observations
(3) estimates sources through an iterative method, thus running ICA multiple times can produce slightly different results

the estimated sources may be opposite signs as the true sources

Outline

- ~~What is source localization?~~
  - ~~sources behind EEG~~
  - ~~smearing & mixing of sources~~
  - ~~what source separation is NOT~~

- ~~Principal Component Analysis (PCA)~~
  - ~~a brief review of linear algebra~~
  - ~~finding a new basis (coordinate system)~~
  - ~~why variance?~~
  - ~~PCA in action~~
  - ~~limitations~~

- ~~Independent Component Analysis (ICA)~~
  - ~~a revised definition of "independence"~~
  - ~~method for finding a new basis~~
  - ~~ICA in action~~
  - ~~limitations~~

- **Other applications of PCA/ICA**
  - **dimensionality reduction (projections)**
  - **"state-space" representation of underlying sources**
  - **classification / clustering**
  - **artifact detection / elimination**

Dimensionality Reduction

By projecting data onto only a few PC's, we reduce the dimensionality of our data.

This is especially useful if we believe the true number of underlying sources is less than the number of channels that we record

$$Y = WX \qquad W = \begin{bmatrix} w_{1,1} & \cdots & w_{1,m} \\ w_{2,1} & \cdots & w_{2,m} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{n,m} \end{bmatrix}$$

Here, we are assuming the first 1 —> k sources in Yr are the "relevant" sources

the n-k sources are "noise".

This isn't always true!

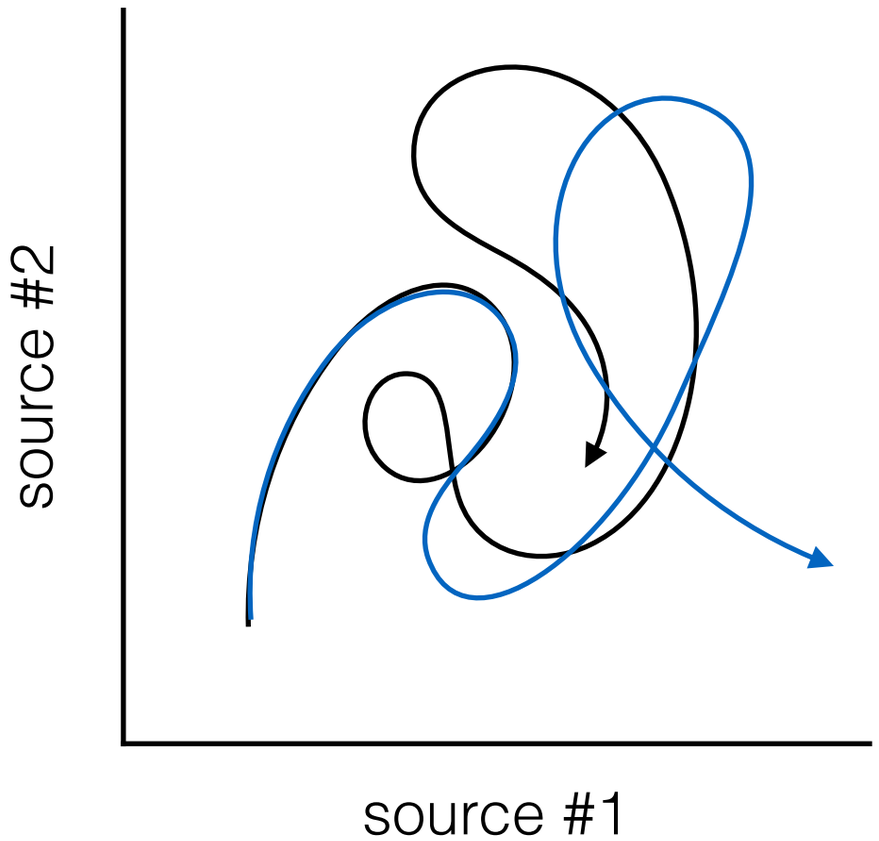
$$Y_r = W_r X \qquad W_r = \begin{bmatrix} w_{1,1} & \cdots & w_{1,k} \\ w_{2,1} & \cdots & w_{2,k} \\ \vdots & \ddots & \vdots \\ w_{m,1} & \cdots & w_{n,k} \end{bmatrix} \qquad k < m$$

“State-space” representations

One question we may have is what the “state” of the brain is during a certain mental task, which may be elusive to traditional analyses (spectral, phase, etc.)

$$Y = WX$$

reconstruct your sources with k ≤ m columns of W



plot the sources against one another

compare “state trajectories” for different conditions

Clustering / classification

Rather than projecting data to recover time-varying sources, we can run PCA/ICA on transposed data.

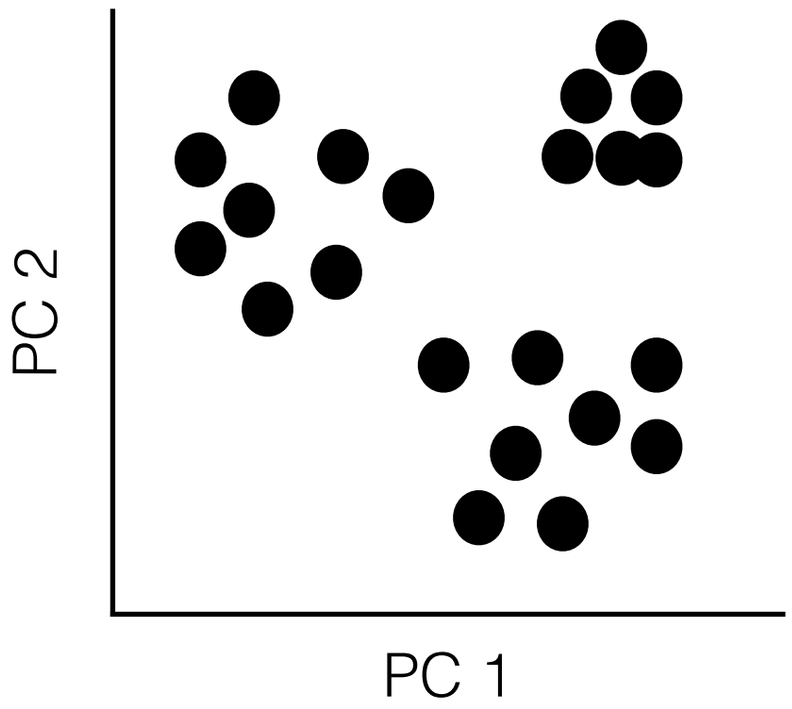
Now, projections tell us the “relatedness” of each time-varying PC to each channel/trial. This can let us cluster and classify different observed time-series

$$W = PCA(X^T)$$

run PCA on transposed data to get time-varying PC “features” found in X

$$Y = WX$$

project onto k ≤ m columns of W



plot the “scores” of each time-varying PC on Y against each other as scatter plots (non-continuous)

Artifact elimination

Both PCA and ICA can be used for artifact elimination.

Because PCA orders PC's according to variance, we can often throw away the last few columns of W to eliminate low-amp. gaussian noise

ICA does a better job at estimating non-gaussian large-amp artifacts like movement and occulomotor

$$Y = WX$$

$$Y_{artifact} = 0$$

$$X_{den} = W^{-1}Y$$

usually manually identify artifacts in Y…