BME I5100: Biomedical Signal Processing

Linear Mixtures and Independent Component Analysis

Lucas C. Parra
Biomedical Engineering Department
City College of New York

Schedule

Week 1: Introduction
Linear, stationary, normal - the stuff biology is **not** made of.

Week 1-4: Linear systems
Impulse response
Moving Average and Auto Regressive filters
Convolution
Discrete Fourier transform and z-transform
Sampling

Week 5-8: Random variables and stochastic processes
Random variables
Moments and Cumulants
Multivariate distributions
Stochastic processes

Week 9-14: Examples of biomedical signal processing
Probabilistic estimation
Harmonic analysis - **estimation** circadian rhythm and speech
Linear discrimination - **detection** of evoked responses in EEG/MEG
Independent components analysis - **analysis** of MEG signals
Auto-regressive model - **estimation** of the spectrum of 'thoughts' in EEG
Matched and Wiener filter - **filtering** in ultrasound

Linear Mixtures – Problem statement

$$\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{S}$$

Q: Given $\boldsymbol{X}$ can you tell what $\boldsymbol{A}$ and $\boldsymbol{S}$ is?

A: Yes! Use prior information on $\boldsymbol{A}$ and $\boldsymbol{S}$.

Linear Mixtures – Problem statement

Basic physics often leads to linear mixing where

- Rows in $\boldsymbol{S}$ are sources $s_i(t)$
- rows in $\boldsymbol{X}$ are sensors readings $x_j(t)$.
- rows in $\boldsymbol{A}$ are the amount the different sources i contribute to a sensor j due to a physical mixing process with coefficients a_{ji}.

Examples are:

$\boldsymbol{X}$	= $\boldsymbol{A}$	* $\boldsymbol{S}$
Acoustic mic. array	= room response	* sound amplitude
Spectroscopy spectra	= concentration	* emission spectra
Hyperspectral image	= abundance	* reflection spectra
EEG electrical potential	= elect. Potential	* electrical coupling
MEG magnetic field	= electrical current	* magnetic coupling

Linear Mixtures – Prior information

Depending which prior information we get different results:

- Columns in ***A*** and rows in ***S*** orthogonal:

  Principal Component Analysis (PCA)

- Rows in ***S*** statistically independent:

  Independent Component Analysis (ICA)

- Rows in S orthogonal and white (or non-stationary):

  Multiple Diagonalization

- ***A*** and ***S*** positive:

  Non-negative Matrix Factorization (NMF)

Assignment 12: Read Lee & Seung Nature 1999, NIPS 1999

Linear Mixtures – Independent Components

Consider an invertible linear mixture

$$\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t) \qquad\qquad \boldsymbol{s}(t) = \boldsymbol{W}\,\boldsymbol{x}(t)$$

where we denoted the inversion by $\boldsymbol{W}=\boldsymbol{A}^{-1}$.

In **Independent Component Analysis** (ICA) one assumes that the sources $\boldsymbol{s}(t)$ are **statistically independent**:

$$p(\boldsymbol{s}(t)) = p(s_1(t), s_2(t), \ldots, s_d(t)) = \prod_{i=1}^{d} p(s_i(t))$$

To estimate the $\boldsymbol{A}$ we use again Maximum Likelihood. The likelihood of i.i.d. observations $x[n]=\boldsymbol{x}(t_n)$, $n=1,\ldots,T$:

$$p(\boldsymbol{x}[1], \ldots, \boldsymbol{x}[T] \| \boldsymbol{A}) = \prod_{n=1}^{T} p(\boldsymbol{x}[n] \| \boldsymbol{A})$$

Linear Mixtures – Independent Components

For any invertible transformation $\boldsymbol{s} = f(\boldsymbol{x})$,

$$p_x(\boldsymbol{x}) = \left|\frac{d\,\boldsymbol{s}}{d\,\boldsymbol{x}}\right| p_s(\boldsymbol{s})$$

In particular for $\boldsymbol{s} = \boldsymbol{A}^{-1}\boldsymbol{x} = \boldsymbol{W}\boldsymbol{x}$

$$p_x(\boldsymbol{x}) = |\boldsymbol{W}|\, p_s(\boldsymbol{s}) = |\boldsymbol{W}|\, p_s(\boldsymbol{W}\,\boldsymbol{x})$$

The log-likelihood is then

$$L(\boldsymbol{W}) = \ln p_X(\boldsymbol{x}[1],\ldots,\boldsymbol{x}[T] \| \boldsymbol{W}) = \sum_{n=1}^{T} \ln p_x(\boldsymbol{x}[n] \| \boldsymbol{W})$$

$$= T \ln|\boldsymbol{W}| + \sum_{n=1}^{T} \ln p_s(\boldsymbol{W}\,\boldsymbol{x}[n])$$

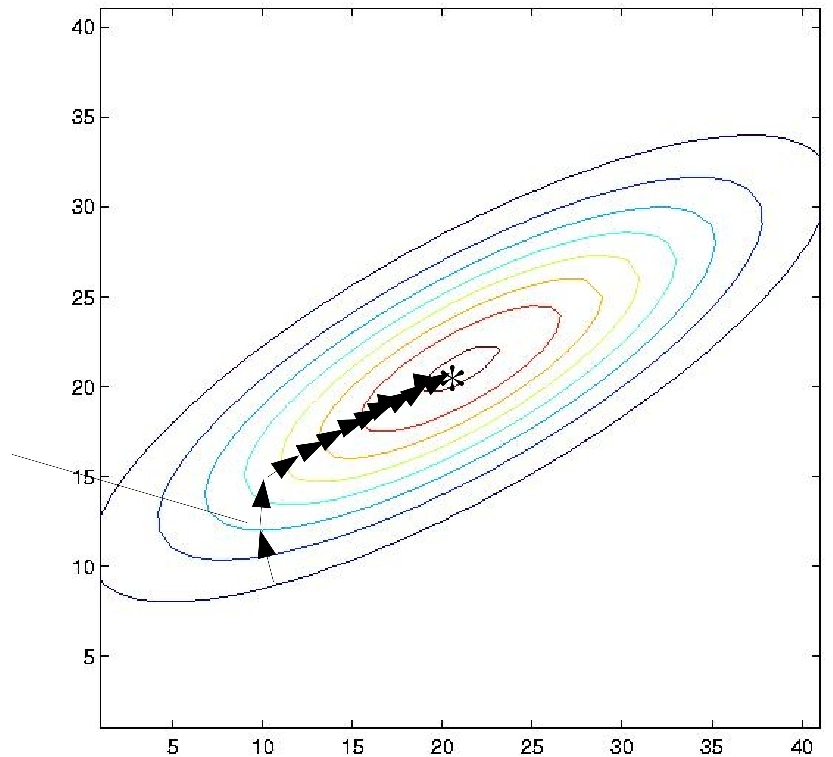
$$= T \ln|\boldsymbol{W}| + \sum_{n=1}^{T} \sum_{i=1}^{d} \ln p_s(\boldsymbol{w}_i^T \boldsymbol{x}[n])$$

Linear Mixtures – Gradient ascent

We can find the maximum of $L(\boldsymbol{W})$ with gradient ascent:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \mu \frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}}$$

$$\frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}}$$



Linear Mixtures – Stochastic gradient ascent

We can find the maximum of $L(\boldsymbol{W})$ with gradient ascent:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \mu \frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{W}}$$

$$= \boldsymbol{W}_t + \mu \frac{\partial}{\partial \boldsymbol{W}} \sum_{n=1}^{T} L(\boldsymbol{x}[n] \| \boldsymbol{W})$$

Stochastic gradient ascent **updates for every sample** n

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \mu \frac{\partial L(\boldsymbol{x}[n] \| \boldsymbol{W})}{\partial \boldsymbol{W}}$$

making the assumption that that instantaneous gradient is a unbiased estimate of the full gradient.

Linear Mixtures – Independent Components

We can find the maximum of $L(\boldsymbol{W})$ with gradient ascent

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \mu\left(\boldsymbol{W}^{-T} + \boldsymbol{u}[n]\boldsymbol{x}^T[n]\right)$$

where we have defined $\boldsymbol{u} = [u_1, \ldots, u_d]^{\mathrm{I}}$ with $u_i = \partial \ln p_s(s_i) / \partial s_i$

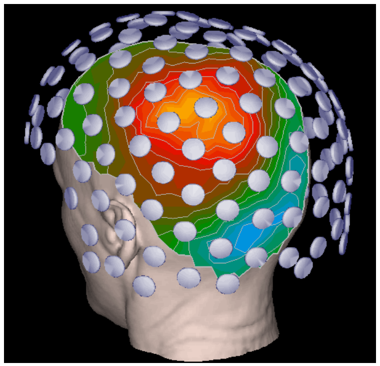
We can always multiplying the gradient with a positive definite matrix, for instance $\boldsymbol{W}^T\boldsymbol{W}$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \mu\left(\boldsymbol{I} + \boldsymbol{u}[n]\boldsymbol{s}^T[n]\right)\boldsymbol{W}$$

If we assume high kurtosis signals (long tails) we can use Laplacian distribution

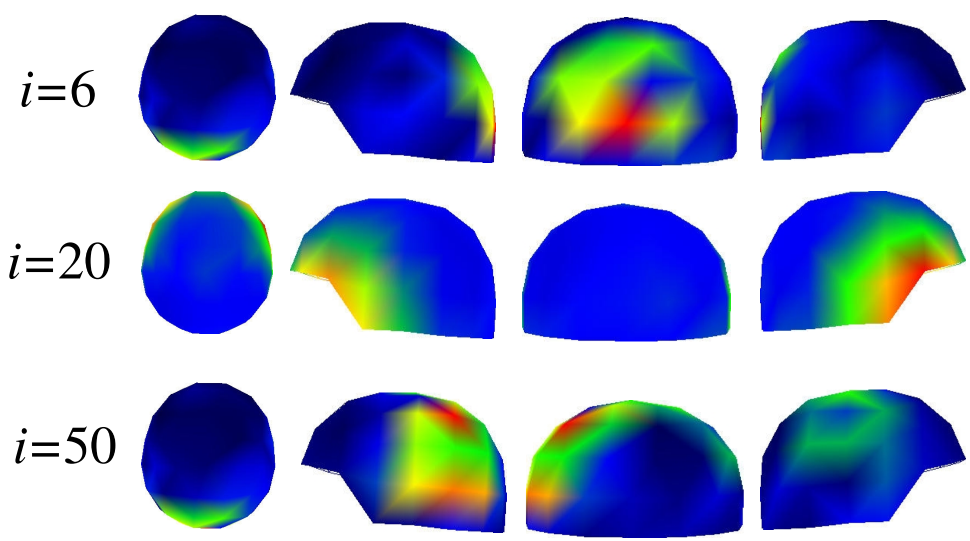
$$p_s(s) = \frac{\lambda}{2}\exp\left(-\lambda\,|s|\right) \qquad u(s) = -\lambda\, sign(s)$$

Linear Mixtures – ICA in MEG

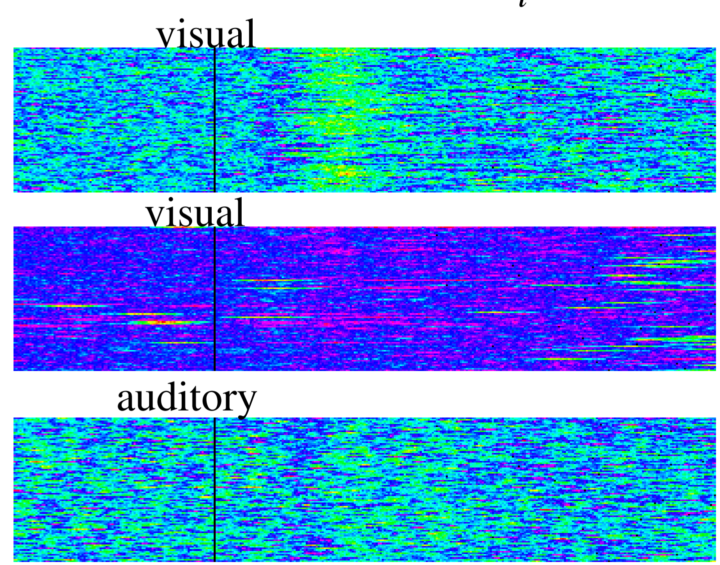


$\boldsymbol{X}$	=	$\boldsymbol{A}$	*	$\boldsymbol{S}$
Magnetic fields measured in SQID sensors		Magnetic coupling or attenuation due to geometry		Effective current flows in neuronal population

i^{th} column in $\boldsymbol{A}$



stimulus locked $s_i(t)$



Linear Mixtures – Principal Components

Maximum Likelihood gives Principal Components if we assume:

- Sources are Gaussian, i.e. $p_s(s) = \sqrt{(2\pi)}\sigma \exp(-\frac{s^2}{2\sigma^2})$
- Mixing are rotations, i.e. $\boldsymbol{W}^{-1} = \boldsymbol{W}^T$

To see this set the gradient of the log-likelihood to zero:

$$0 = T\boldsymbol{W}^{-T} + \sum_{n=1}^{T} \boldsymbol{u}[n]\boldsymbol{x}^T[n]$$

For a Gaussian, $u = \partial \ln p(s)/\partial s = -s/\sigma^2$. With $\boldsymbol{\Lambda} = diag(\sigma_1^2, \ldots, \sigma_d^2)$

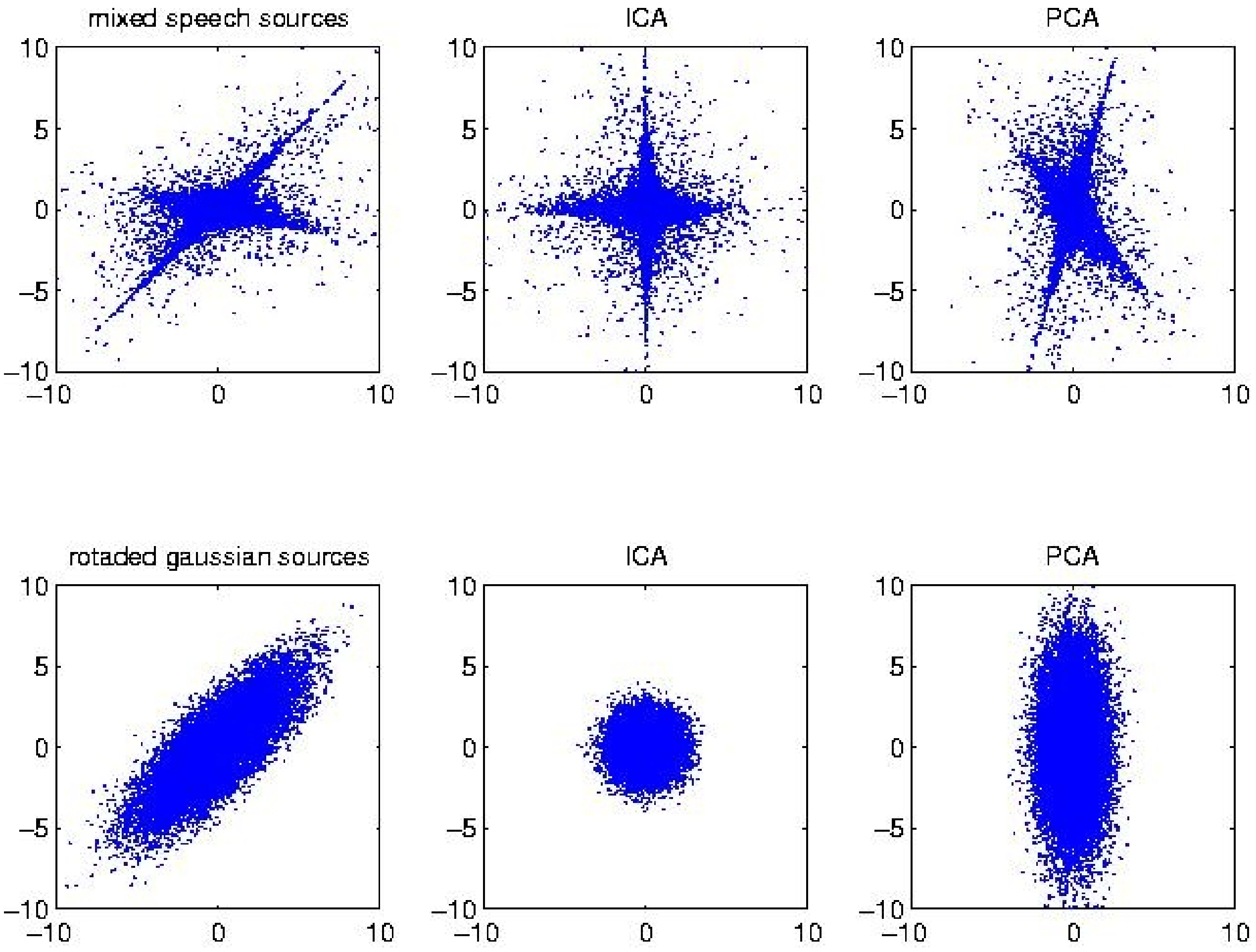
$$\boldsymbol{W}^{-T} = \frac{1}{T}\sum_{n=1}^{T} \Lambda^{-1}\boldsymbol{s}[n]\boldsymbol{x}^T[n] = \frac{1}{T}\sum_{n=1}^{T} \Lambda^{-1}\boldsymbol{W}\boldsymbol{x}[n]\boldsymbol{x}^T[n] = \Lambda^{-1}\boldsymbol{W}\boldsymbol{R}_x$$

Using orthogonality we obtain PCA:

$$\boldsymbol{R}_x = \boldsymbol{W}^T \Lambda \boldsymbol{W}$$

Linear Mixtures – ICA and PCA

Comparing results of ICA and PCA



Linear Mixtures – Multiple diagonalization

Statistical independence implies for all $i \neq j, t, l, n, m$:

$$E[s_i^n(t)\, s_j^m(t+l)] = E[s_i^n(t)]\, E[s_j^m(t+l)]$$

So far we have talked about same number of sensors than sources. In general for M sources and N sensors each t, l, n, m gives $M(M-1)/2$ conditions for NM unknowns in $\boldsymbol{A}$.

Sufficient conditions if we use multiple:

use	sources assumed	resulting algorithm
n, m	non-Gaussian	ICA
t	non-stationary	multiple decorrelation
l	non-white	multiple decorrelation

Linear Mixtures – Multiple diagonalization

For **stationary non-white sources** we have:

$$\boldsymbol{R}_s(l) = E\left[\boldsymbol{s}(t)\,\boldsymbol{s}^T(t+l)\right]$$

Which is diagonal assuming second order independent sources.
The measured cross-correlation $\boldsymbol{R}_x(l) = E[\boldsymbol{x}(t)\,\boldsymbol{x}^T(t+l)]$ is then

$$\boldsymbol{R}_x(l) = \boldsymbol{A}\,R_s(l)\,\boldsymbol{A}^T$$

Combing these equations for two time delays $l=l_1,l_2$ leads to a generalized eigenvalue problem for $\boldsymbol{A}$,

$$\boldsymbol{R}_x(l_1)\,\boldsymbol{R}_x^{-1}(l_2)\,\boldsymbol{A} = \boldsymbol{A}\,\Lambda_s(l_1)\,\Lambda_s^{-1}(l_2)$$

Warning: Generalize Eigenvalue not robust to noise!
For increased stability diagonalize multiple delays l. This is known as the SOBI algorithms (second order blind identification)

Linear Mixtures – Multiple diagonalization

For **non-stationary white sources** we have:

$$\boldsymbol{R}_s(t) = E[\boldsymbol{s}(t)\,\boldsymbol{s}^T(t)]$$

Which is diagonal assuming second order independent sources.
The measured cross-correlation $\boldsymbol{R}_x(t) = E[\boldsymbol{x}(t)\,\boldsymbol{x}^T(t+l)]$ is then

$$\boldsymbol{R}_x(t) = \boldsymbol{A}\,R_s(t)\,\boldsymbol{A}^T$$

Combing these equations for two time intervals $t=t_1,t_2$ leads to a generalized eigenvalue problem for $\boldsymbol{A}$,

$$\boldsymbol{R}_x(t_1)\,\boldsymbol{R}_x^{-1}(t_2)\,\boldsymbol{A} = \boldsymbol{A}\,\Lambda_s(t_1)\,\Lambda_s^{-1}(t_2)$$

This is equivalent to “Common Spatial Pattern”, which diagonalizes two covariance matrices measured at different times)
Warning: Generalize Eigenvalue not robust to noise!

Linear Mixtures – Multiple diagonalization

For **non-Gaussian sources** we have:

$$E[s_i^u s_j^v] = E[s_i^u] E[s_j^v] \quad , \quad i \neq j$$

From this one can derive for a linear combination of 4-order tensors (cross-cumulants) with symmetric matrix ***M***

$$\boldsymbol{C_s}(\boldsymbol{M}) = E[\boldsymbol{s}^T \boldsymbol{M}\, \boldsymbol{s}\, \boldsymbol{s}\, \boldsymbol{s}^T] - \boldsymbol{R_s}\, Trace(\boldsymbol{M}\, \boldsymbol{R_s}) - 2\, \boldsymbol{R_s}\, \boldsymbol{M}\, \boldsymbol{R_s}$$

The following diagonalization condition

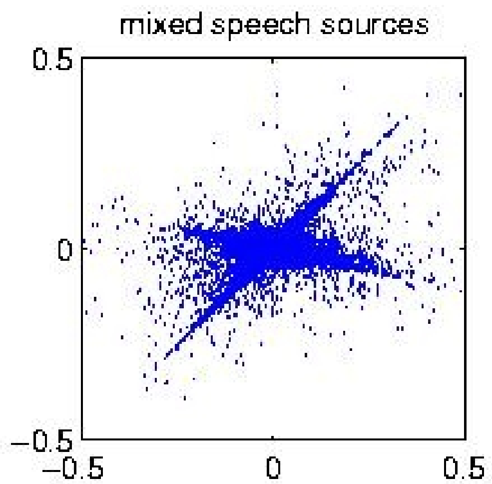
$$\boldsymbol{C_x} = \boldsymbol{A}\, C_s(\boldsymbol{A}^T \boldsymbol{A})\, \boldsymbol{A}^T$$

This can again - in combination with diagonal covariance - be combined to a generalized eigenvalue equation.

And again, for stability one should use use more than two cumulants, which leads to the “JADE” algorithms.

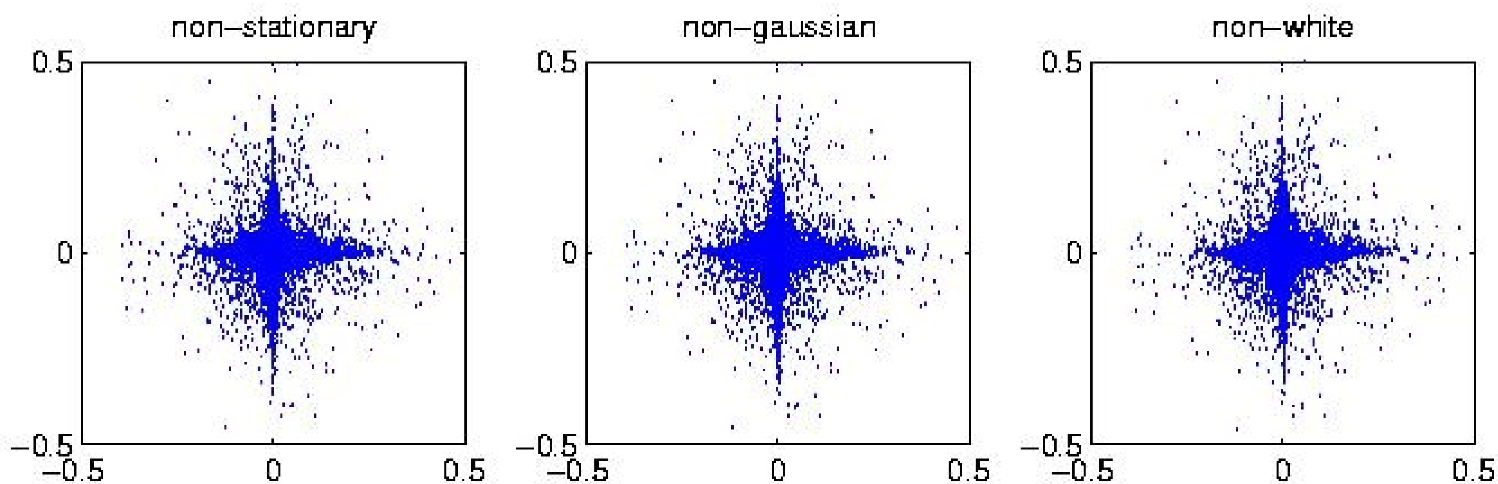
Linear Mixtures – Multiple diagonalization

Comparing different diagonalization criteria



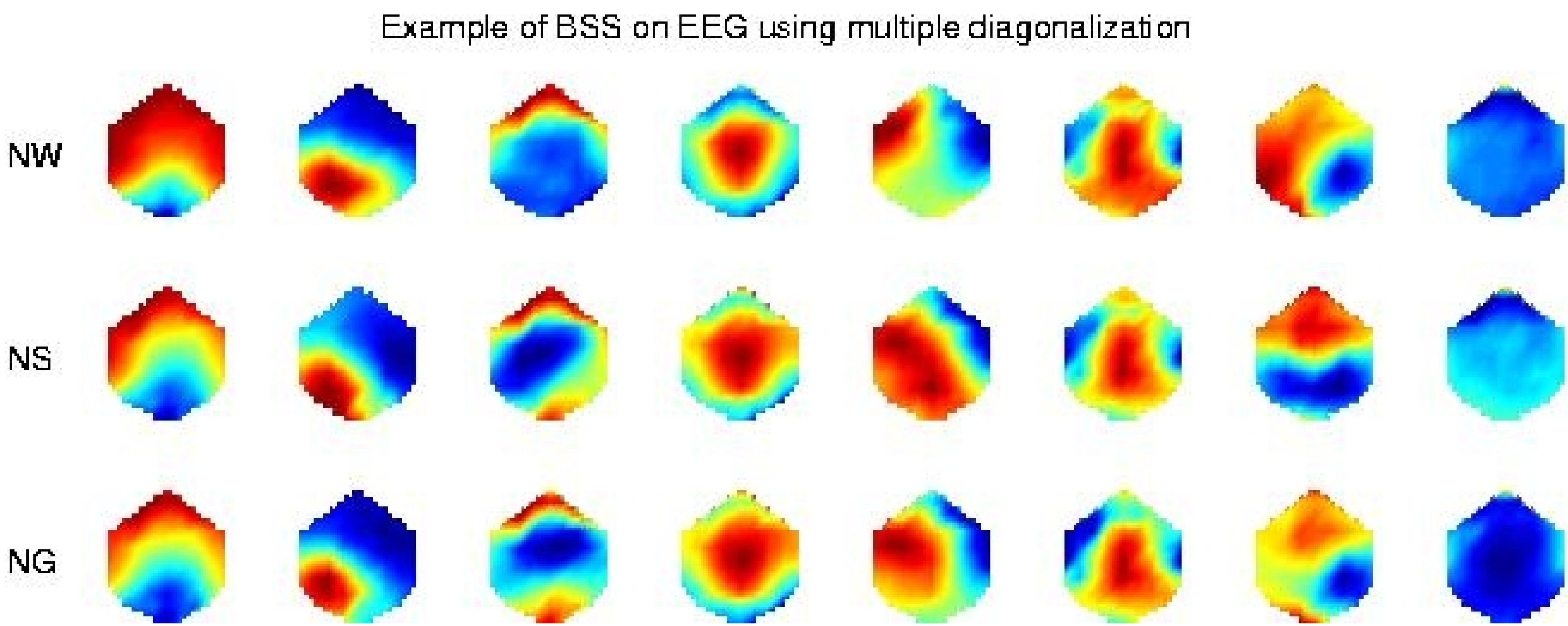

```
% linear mix of sourses S
X=A*S;

% Separation based on Generalized Eigenvalues
[W,D]=eig(X*X',Q);
S=W' *X;
```

Results with Q assuming:



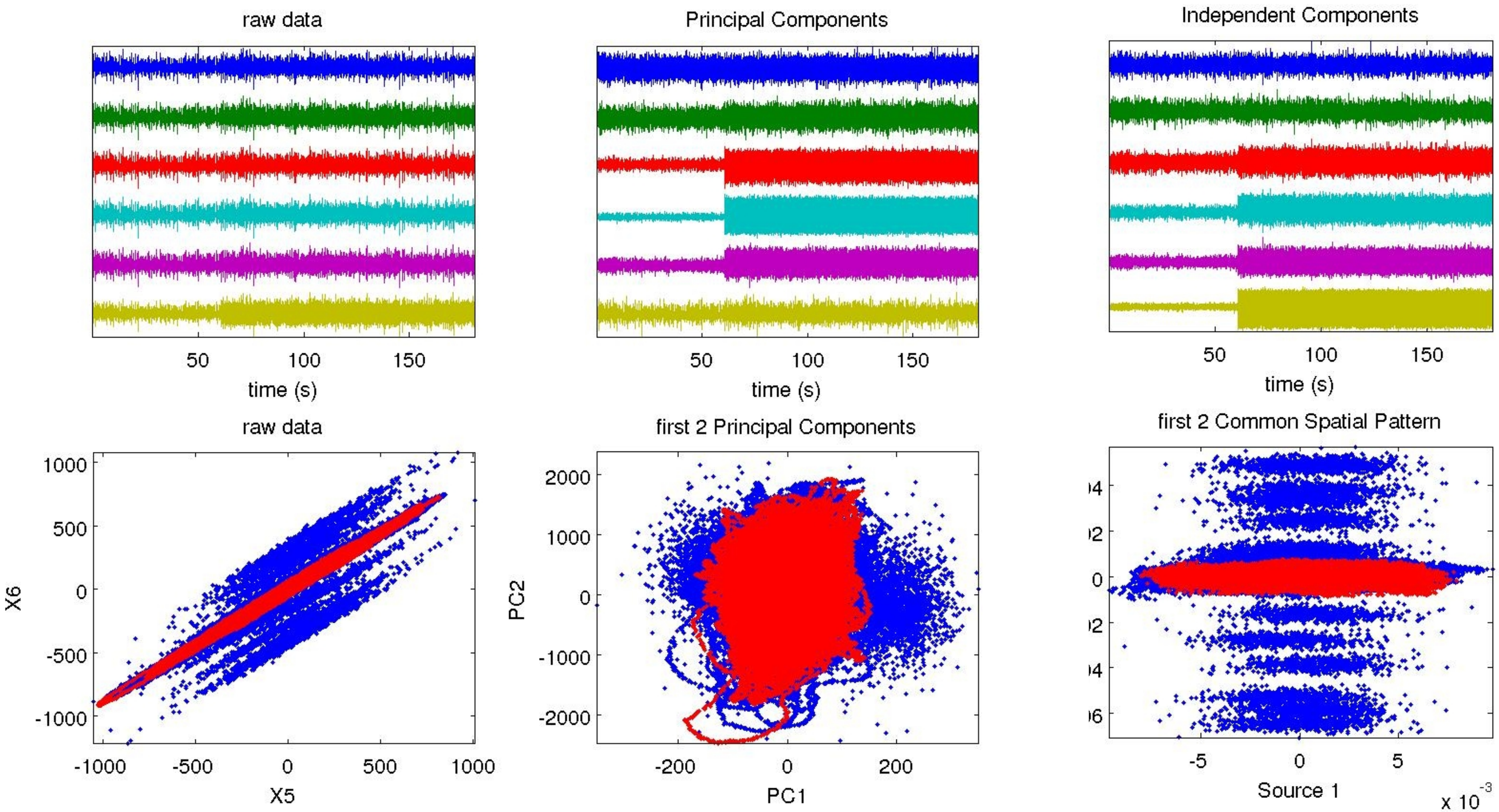
Linear Mixtures – Multiple diagonalization

X	=	A	*	S
Electrical potentials on the skull surface		Electric coupling or attenuation due to tissue resistance		Large scale potential of neuronal population



Linear Mixtures – Multiple diagonalization

Example: 6D Local Field Potentials




```
R=X(t2,:)*X(t2,:)';
Q=X(t1,:)*X(t1,:)';
```

```
[W,D]=eig(R);
```

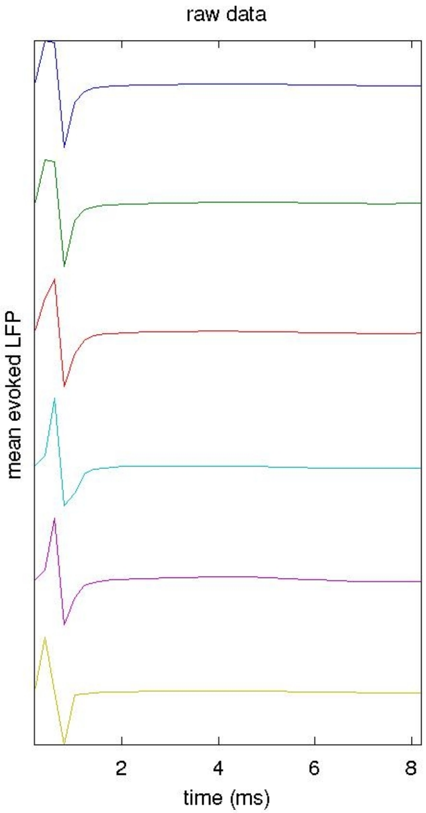
```
[W,D]=eig(R,Q);
```

```
S=W'*X;
```

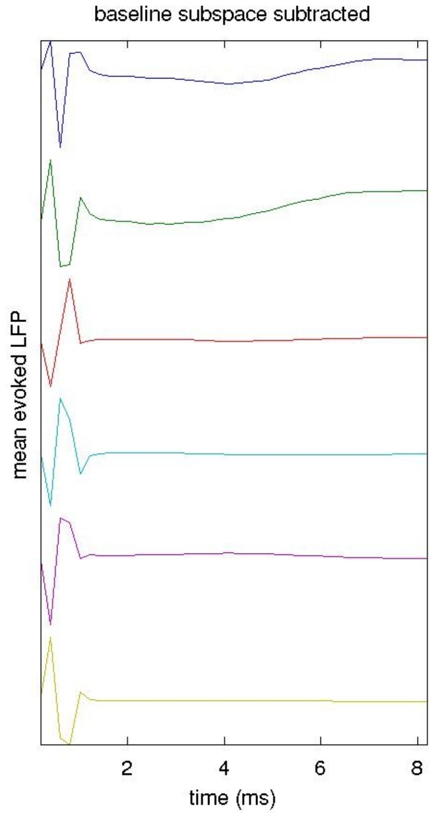
Linear Mixtures – Multiple diagonalization

Stimulus triggered evoked responses of these LFP (averaged over multiple repeats) for different spatial projections of the data: $Y = W^T X$, with baseline covariance Q and stimulus covariance R:

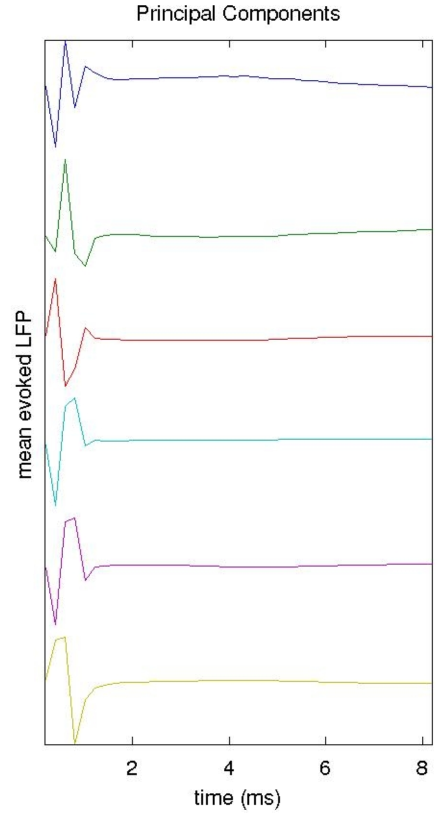
```
W=eye(6)
```



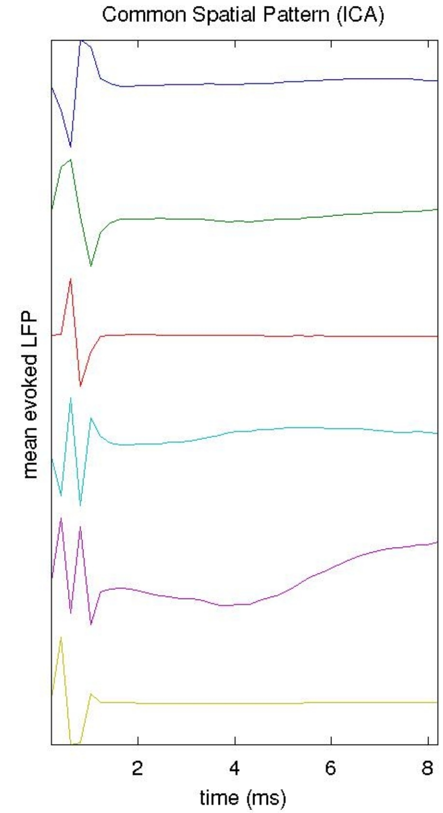

```
[U,D]=eig(Q);
W=(I-U(:,6)*U(:,6));
```




```
[U,D]=eig(R);
W=U;
```




```
[U,D]=eig(R,Q);
W=U;
```



Linear Mixtures – Maximum SNR component

Assume we are looking for a linear response $r(t)$

$$r(t) = \boldsymbol{w}^T \boldsymbol{x}(t)$$

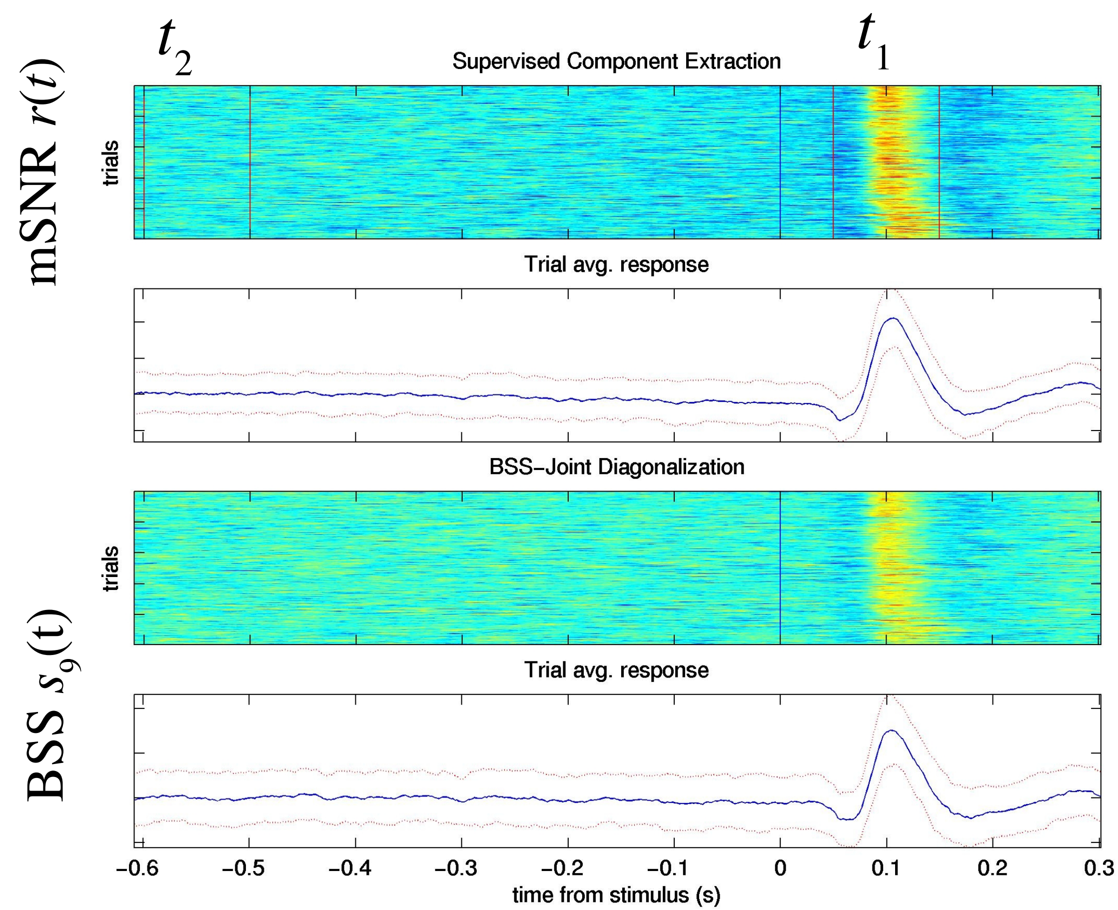
which has **maximal power** during specific times t_1 as compared to the **power** of baseline activity during times t_2. That is, we are looking for a linear component $\boldsymbol{w}$ that maximize the power ratio or signal to noise ratio (SNR)

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{argmax} \frac{E[r^2(t_1)]}{E[r^2(t_2)]} = \underset{\boldsymbol{w}}{argmax} \frac{\boldsymbol{w}^T \boldsymbol{R}_{\boldsymbol{x}}(t_1) \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{R}_{\boldsymbol{x}}(t_2) \boldsymbol{w}}$$

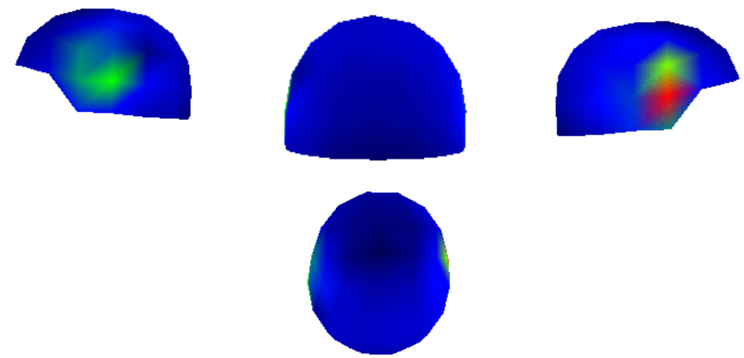
With the usual definition, $\boldsymbol{R}_{\boldsymbol{x}}(t) = E[\boldsymbol{x}(t)\, \boldsymbol{x}^T(t)]$. The solution if given by the maximum generalized eigenvector. Hence, the first component in the previous approach with two stationarity times has maximum SNR.

Linear Mixtures – Maximum SNR component

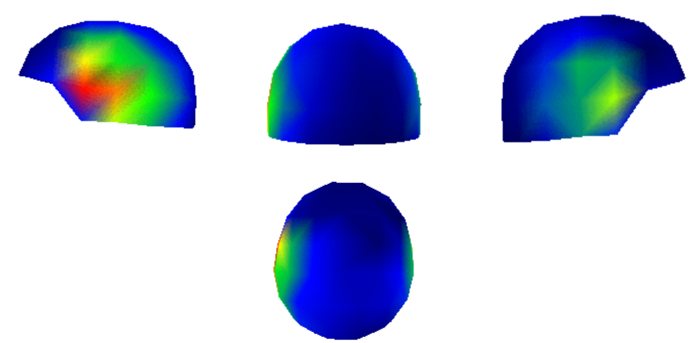
Comparing BSS based on non-stationarity and non-white in MEG



"inverse" of ***w***



9th column in ***A***



auditory locked

Assignment: source separation

Use PCA, and multiple diagonalization to generate various projections of EEG data.

1. Load the file `eeg-vep.mat` from the class webpage. It should have epoched data with 160 channels, 545 samples and 116 trials saved in variable `eeg`.
2. Stack up the trials so that you have a data matrix of channels by samples (160 x 1630).
3. Compute the covariance and from this principal components of this data. Show the eigenvalue spectrum (on a dB scale). Display first principal component (the eigenvector with the strongest eigenvalue) on the scalp using the `topoplot()` function using the corresponding location file provided on the website.
4. Project the data onto the first two principal component and display the result as an images of trials by samples (116x545 for each of the two components). Also show the average across trials for the two components (two curved using plot).
5. Repeat steps 2 trough 4 using instead the first and second half of the samples (samples 1:272 and 273:545) to obtain two covariance matrices. Find the eigenvectors that diagonalize both these matrices (generalized eigenvectors).
6. Repeat step 5 but now using the covariance of all the data and as the second correlation matrix use the the cross-correlation of the data with a version of the data that is delayed by K samples (try different delays). Be sure to symmetrize this matrix with Rxy=0.5*(Rxy+Rxy');

In total you should have 3 figures with results for 3 different techniques (steps 4, 5 and 6). For each method combine plots with subplot into single figure. Label all axis and use milliseconds on the time axis.