Solutions 10

Solution 1. Using the Karhunen-Loève Transform in Matlab

(a) In Matlab:

```
N = 5; x = randn(N,1);
```

(b) First, let us generate the Gaussian normalized sequence $\mathbf{X}[n]$ of length M:

```
M = 10000; x = randn(N,M);
```

Now, choose the matrix $\mathbf{A}$, as $\mathbf{A} = \mathbf{V}_Y \cdot \mathbf{\Lambda}_Y^{1/2}$, where $\mathbf{V}_Y$ and $\mathbf{\Lambda}_Y$ are eigenvectors and eigenvalues of the autocorrelation matrix $\mathbf{R}_Y$.

```
Ry = [1.9 0.5 0.3 0.2 0.05;
      0.5 2.3 0.4 0.2 0.1;
      0.3 0.4 1.5 0.9 0.7;
      0.2 0.2 0.9 1.1 0.8;
      0.05 0.1 0.7 0.8 1.2];
[Vy,Ly] = eig(Ry); A = Vy*Ly^0.5; y = A*x;
```

(c) The correlation of the generated sequence $\mathbf{y}$ is evaluated by

```
Ry1 = (y*y')/M;
```

The correlation $\hat{\mathbf{R}}_y$ approximates the expected correlation $\mathbf{R}_y$. They are not exactly the same because of the finite length of the Gaussian sequence $\mathbf{x}$. As the length M grows, the approximated correlation $\hat{\mathbf{R}}_y$ is closer to the expected correlation $\mathbf{R}_y$.

(d) The KLT matrix $\mathbf{T}$ contains the eigenvectors of the estimated correlation matrix $\hat{\mathbf{R}}_y$ as rows in descending order of the corresponding eigenvalues. The KLT and the correlation $\mathbf{R}_z$ are calculated by

```
% Compute the eigenvectors and eigenvalues of the estimated correlation matrix Ry1
[Vy1,Ly1] = eig(Ry1);
% Then, sort the eigenvalues
[Lsorted,I] = sort(diag(Ly1));
% Arrange them in descending order (sort gives ascending order)
I = I(length(I):-1:1);
% Take the corresponding columns from Vy1 and put them as rows in T
T = Vy1(1:N,I)';
% Apply the KLT
z = T*y;
% Compute the correlation Rz
Rz = (z*z')/M;
```

The correlation $\mathbf{R}_z$ is diagonal. This is expected since we used the estimated correlation matrix $\hat{\mathbf{R}}_y$. The KLT is obviously signal-dependent, because it is constructed using the properties of the generated signal.

Solution 2. KARHUNEN-LOÉVE TRANSFORM

Since $x[n]$ is periodic, the correlation function is also periodic, i.e., $R_x[n] = R_x[n+kN], \forall n, k \in \mathbb{Z}$. Therefore, if we use blocks of N consecutive samples of $x[n]$, we obtain vectors whose correlation matrix is circulant.

(a) In this case $N = 2$ and $R_x[n] = [1, 0.5]$ and the correlation matrix is

$$\mathbf{R_x} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},$$

and the KLT is the matrix

$$\mathbf{H} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Remark that this is also the DFT of size 2 (properly scaled).

(b) In this case, we should use the DFT of size 4 and normalize $\sqrt{4}$, i.e.,

$$\mathbf{H} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \end{bmatrix}$$

we have that $\mathbf{Y} = \mathbf{H}^\star\mathbf{X}$; therefore,

$$\mathbf{R_Y} = \mathbf{H}^\star\mathbf{R_x}\mathbf{H} = \begin{bmatrix} 2 & & & \\ & 0.8 & & \\ & & 0.4 & \\ & & & 0.8 \end{bmatrix}$$

Solution 3. SPIKE SORTING

1) We have a WSS process that takes continuous values corrupted by a white noise. Among the denoising methods we have seen in class, the optimal one is the **Wiener filter**.

We need the signal $X_0[n]$ to be WSS and also $X_0[n]$ and $X[n]$ to be jointly WSS. The first condition is given as assumption and the second satisfied since $X[n]$ is the sum of $X_0[n]$ and another WSS process $W[n]$ independent of $X_0[n]$.

In this specific case, the transmittance $H(e^{j\omega})$ of the Wiener filter $h[n]$ is given by the expression

$$H(e^{j\omega}) = \frac{S_{X_0X}(\omega)}{S_X(\omega)} = \frac{S_{X_0}(\omega)}{S_X(\omega)} = \frac{S_X(\omega) - \sigma_W^2}{S_X(\omega)},$$

where $\sigma_W^2 = 4$.

Notice that since we do not know $X_0[n]$ we cannot directly compute its power spectral density $S_{X_0}(\omega)$ but rather write it as $S_X(\omega) - \sigma_W^2$.

In order to apply the Wiener filter approach we need to

- Estimate $S_X(\omega)$ from the samples $x[1], \ldots, x[500000]$ for instance using the Periodogram

$$\mathbf{R}_X = \mathbf{x}[n] \star \overline{\mathbf{x}}[-n]$$
$$\widehat{S}_X(\omega) = \mathcal{F}\{\mathbf{R}_X\}(\omega)$$

- Compute the transmittance of the filter

$$H(e^{j\omega}) = \frac{\widehat{S}_X(\omega) - 4}{\widehat{S}_X(\omega)}$$

- Compute the Discrete Time Fourier Transform - DTFT of the samples $x[1], \ldots, x[500000]$

$$\chi(e^{j\omega}) = \sum_{n=1}^{500000} x[n] e^{-j\omega n}\,.$$

- Compute the denoised signal as the inverse DTFT of

$$\widehat{x}_0[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \chi(e^{j\omega}) H(e^{j\omega}) d\omega\,.$$

Notice that here we can also consider the Discrete Fourier Transform DFT for both computing the Periodogram and the convolution as inverse transformation of the product.

2) The dimension of the data is the number of identified shapes, that is M=5001, while the variables are the N=6 characteristics.

In order to be able to apply PCA, the variable vector must be WSS (verified since the signal itself is supposed WSS) and centred.

We call $\boldsymbol{c}_m = [c_m[1], \ldots, c_m[6]]$ the variables of the problem, that is, amplitude, energy, duration, median, mode, and mean of the pulse shape, where $m = 1, \ldots, 5001$ indicates the pulse shape.

- Center the variable vector $c_m[1], \ldots, c_m[6]$, for every $m = 1, \ldots, 5001$.
- Compute the empirical covariance Matrix

$$\widehat{\mathbf{R}}_{\boldsymbol{c}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{c}_m * \boldsymbol{c}_m^H = \frac{1}{M} \boldsymbol{C} * \boldsymbol{C}^H,\ (N \times N)\,,$$

  Where, $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{c}_1 & \ldots & \boldsymbol{c}_M \end{bmatrix}, (N \times M)$ with N=6 and M=5001.

- Compute the unitary matrix of eigenvectors of $\widehat{\mathbf{R}}_{\boldsymbol{c}}$ and the corresponding eigenvalues $\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$. We assume here the eigenvalues to be ordered from the highest to the lowest value (the eigenvectors are also ordered according to the corresponding eigenvalue).

- Identify the eigenvalues that account for most of the total sample variation, say $\lambda_1, \ldots, \lambda_k$.
- Compute the matrix of the principal components

$$\boldsymbol{Z} = \boldsymbol{V}^T \boldsymbol{Y}\,, (N \times M)\,, \quad \text{with with N=6 rows and M=5001 columns}\,.$$

  The eigenvalues represents the variance of the principal components.
- Extract from $\boldsymbol{Z}$ the first k rows which correspond to the k eigenvalues accounting for most of the total sample variation. The so obtained k principal components $z_m[1], \ldots, z_m[k]$, $m = 1, \ldots, 5001$, represent the essential characteristics of the data.

3) We have 3 clusters in a 2-dimensional data plot.

   We can model each cluster as being the sampling of a 2-dimensional Gaussian distribution. The overall 2-dimensional data can therefore been seen as a mixture of three 2-dimensional Gaussian distributions.

   The parameters of the Gaussian distributions are their means and variances, and their estimation provide the central point and the variance of each cluster.

   For simplicity, we assume the mixture model to be i.i.d.

   - Define the model.
     For a couple of values $\boldsymbol{z} = [z[1]z[2]]$ of the two principal components the model reads

$$= \sum_{k=1}^{3} \pi_k \exp\left(-\frac{1}{2}((\boldsymbol{z} - \mathbf{m}_k)^t \boldsymbol{C}_k^{-1}(\boldsymbol{z} - \mathbf{m}_k)\right).$$

     where, due to the uncorrelation of the principal components,

$$\boldsymbol{C}_k = \text{diag}((\sigma_1^2[k], \sigma_2^2[k])\,.$$

     When considering all the data, that is the 5001 couples of values of the principal components, having supposed the mixture model to be i.i.d. we obtain

$$= \prod_{m=1}^{5001} \sum_{k=1}^{3} \pi_k \exp\left(-\frac{1}{2}((\boldsymbol{z}_m - \mathbf{m}_k)^t \boldsymbol{C}_k^{-1}(\boldsymbol{z}_m - \mathbf{m}_k)\right).$$

   - Feed the data (5001 values of the couples $(z_m[1], z_m[2])$, $m = 1, \ldots, 5001$ into an EM algorithm for a 3 component mixture model with diagonal covariance
   - The estimated $\mathbf{m}_k$ and $\boldsymbol{C}_k$ provides the centre of the three clusters and their variance.