Statistical Signal Processing
Final Exam

Thursday, 22 June 2017

You will hand in this sheet together with your solutions.

Write your personal data (please make it readable!).

Seat Number:

Family Name:

Name:

Read Me First!

You are allowed to use:

- A handwritten cheatsheet (2 A4 sheets, double sided) summarizing the most important formulas (no exercise text or exercise solutions);
- A pocket calculator.

You are definitively not allowed to use:

- Any kind of support not mentioned above;
- Your neighbor; Any kind of communication systems (smartphones etc.) or laptops;
- Printed material; Text and Solutions of exercises/problems; Lecture notes or slides.

Write solutions on separate sheets, *i.e.* no more than one solution per paper sheet.

Return your sheets ordered according to problem (solution) numbering.

Return the text of the exam.

Warmup Exercises

This is a warm up problem .. do not spend too much time on it. Please provide justified, rigorous, and simple answers. If needed, you can add assumptions to the problem setup.

Exercise 1. CORRELATION (4 PTS)
We have recorded $N = 100000$ samples $x[1], \ldots, x[100000]$ of a w.s.s. signal. Using the empirical mean $\widehat{m}$, we observe that the mean of the signal is **clearly not zero**, *i.e.*, $|\widehat{m}| \gg 0$.

We would like to compute the empirical correlation $\widehat{R}_X[k]$ for $k = 0, 1, \ldots, 5$.

Which form of the empirical correlation should we use? (*Only one answer is correct, and you have to justify it precisely*)

- Absolutely only the unbiased correlation!
- Absolutely only the biased correlation!
- Either one of the two, it does not make a relevant difference!

Write the chosen expression(s) of the empirical correlation, given the samples $x[1], \ldots, x[100000]$ (both if you choose "either one of the two").

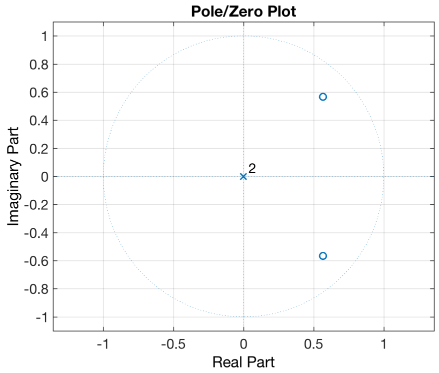
Solution 1.

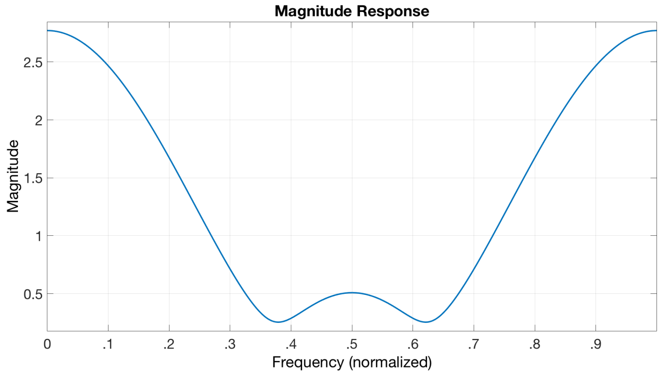
- Biased correlation $\widehat{R}_b(k) = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N-k} x[n+k]x[n]^*$;
- Unbiased correlation $\widehat{R}_u(k) = \dfrac{1}{N-k} \displaystyle\sum_{n=1}^{N-k} x[n+k]x[n]^*$.

The two equations hold for $k \geq 0$, and we set $\widehat{R}(-k) = \widehat{R}_b^*(k)$.
Give that $N = 100000$ and $k = 0, 1, \ldots, 5$, we have $\dfrac{1}{N} \approx \dfrac{1}{N-k}$, consequently the two estimates of the correlation provide values that do not preset a relevant difference.

Exercise 2. A SIMPLE SYSTEM (4 PTS)
Below we have two plots:

- A z-plane plot (on the left) with two zeros with magnitude $a = 0.8$ and phase $\varphi = \pm\pi/4$, and two poles in the origin (recall that this is a z-plane and not a z^{-1}-plane!);
- The plot of the magnitude of a transfer function (on the right).





1) Write the z transform $H(z)$ corresponding to the z-plane plot (on the left).

2) Compute the impulse response $h(n)$ corresponding to the z transform $H(z)$.

3) Does the plot on the right correspond to the magnitude of the transfer function associated to $H(z)$? Justify precisely your answer.

Solution 2.

1) The two zeros are $z_1 = 0.8e^{j\pi/4}$ and $z_2 = 0.8e^{-j\pi/4}$, and the two poles $p_1 = p_2 = 0$. Therefore the corresponding z-transforms reads

$$\begin{aligned} H(z) &= \frac{(z-z_1)(z-z_2)}{(z-p_1)(z-p_2)} = \frac{(z-z_1)(z-z_2)}{z^2} = (1-z_1z^{-1})(1-z_2z^{-1}) \\ &= 1-(z_1+z_2)z^{-1} + z_1z_2z^{-2} = 1-1.6\cos(\pi/4)z^{-1} + (0.8)^2z^{-2} \\ &= 1-0.8\sqrt{2}z^{-1} + (0.8)^2z^{-2}\,. \end{aligned}$$

2) From the z-transform $H(z) = 1-0.8\sqrt{2}z^{-1} + (0.8)^2z^{-2}$ the impulse response is

$$h(0) = 1\,,\ h(1) = -0.8\sqrt{2}\,,\ h(2) = (0.8)^2\,,\ h(k) = 0\ \forall k \neq 1,2,3\,.$$

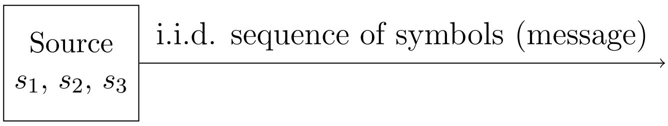
3) The module of the transfer function shows a minimum in the high frequencies $[0.25, 0.75]$ (in radiants $[\frac{\pi}{2}, \frac{\pi 3}{2}]$) and maximum in the low frequencies $[0, 0.25] \cup [0.75, 1]$ (in radiants $[0, \frac{\pi}{2},] \cup [\frac{\pi 3}{2}, 2\pi]$) while the zeros of the z-plane plot are in the low frequencies. Therefore, the two plots do no belong to the same filter.

Main Problems

Here comes the core part of the exam .. take time to read the introduction and each problem statement. Please provide justified, rigorous, and simple answers. Remember that you are not simply asked to describe statistical signal processing tools, but you are rather asked to describe how to apply such tools to the given problem. If needed, you can add assumptions to the problem setup.

Exercise 3. A Professional Transmission System (40 pts)

Consider a source generating symbols with three possible outcomes, s_1, s_2, and s_3, with equal probability. The source generates a sequence of $N = 1000$ i.i.d. symbols over time, that is, a logical message that can be modeled as an i.i.d. process.



In order to transmit the symbols, they are first converted into a physical message (a signal!) and then send over a channel.

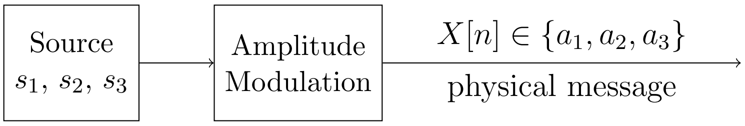
The available channel is an additive Gaussian channel (that is, Gaussian white noise is added to the physical message during the transmission over the channel).

We would like to analyze different transmission systems and you are asked to performs such analysis, as described in the following.

Remark: Part A) and B) are independent!

A) Amplitude Modulation

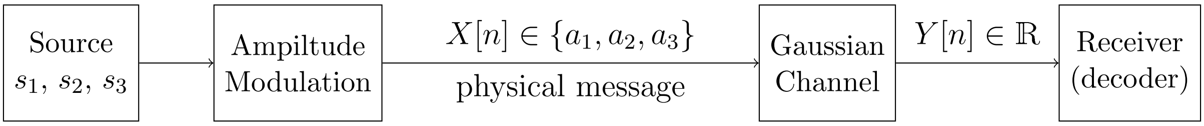
Each possible source outcome, s_1, s_2, and s_3, is converted into a numerical value (*e.g.*, a voltage level), a_1, a_2 and a_3, respectively. Consequently the logical message is converted into a a physical message of $N = 1000$ samples, modeled as stochastic process $X[1], \dots, X[1000]$, with outcome $x[1], \dots, x[1000]$, where each signal sample takes one of three possible values a_1, a_2 and a_3.



At the receiver side (output of the Gaussian channel) we have

$$Y[n] = X[n] + W[n]\,, \quad n = 1, \dots, 1000\,,$$

where $W[n]$ is a white Gaussian noise, centered, with variance σ_W^2.



By defining $\boldsymbol{Y} = [Y[1], \ldots, Y[1000]]$, and the corresponding values $\boldsymbol{y} = [y[1], \ldots, y[1000]]$, and by considering that the number of the possible outcomes of source symbols is known ($K = 3$), let's find the probabilistic model for the received signal, that is:

A.1) Write the cumulative distribution function $F_{\boldsymbol{Y}}(\boldsymbol{y}) = P(\boldsymbol{Y} \leq \boldsymbol{y})$ of the received message;

A.2) Write the corresponding probability density function $f_{\boldsymbol{Y}}(\boldsymbol{y})$.

Due to the noise $W[n]$, the received signal $Y[n]$ cannot be directly decoded. We must first de-noise it!

A.3) Propose a de-noising method by precisely justifying you choice (WITHOUT describing the method in details);

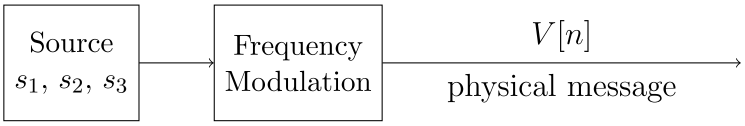
A.4) Given the received signal samples $(y[1], \ldots, y[1000])$ and the number of source symbols ($K = 3$), describe in details the proposed method. Write every step of your method (as a bullet list) clearly indicating the input and output of each step.

B) Frequency Modulation

Each possible source outcome, s_1, s_2, and s_3, is converted into a real sinusoid with (normalized) frequency f_1, f_2, and f_3, respectively. More precisely:

- Each symbol of the logical message corresponds to a sinusoid of 100 samples;
- The (normalized) frequencies are $f_1 = 0.0440$, $f_2 = 0.0494$, and $f_3 = 0.0523$;
- The numerical value of the sinusoid amplitude is 2.

Given that the source generates a sequence of $N = 1000$ symbols, and that each symbol is transformed into a real sinusoid of 100 samples, the corresponding physical message consists of $M = 1000 \times 100 = 100000$ samples, modeled as a stochastic process $V[1], \ldots, V[100000]$, with outcome $v[1], \ldots, v[100000]$.



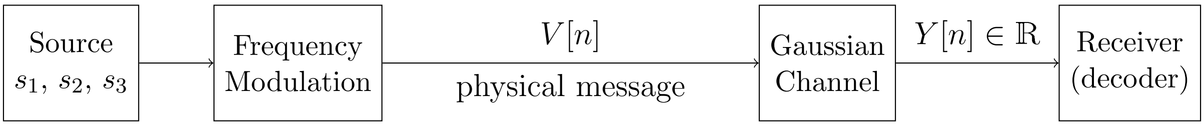
The signal $V[n]$ is a very particular type of harmonic process, since for each segment of 100 samples we have only one sinusoid and not the sum of three sinusoids. A quick Matlab check can show you that for a large number of segments of 100 samples (here we have 1000 of them!) we can approximate $V[n]$ as a standard harmonic process, that is, as the sum of three sinusoids, with frequencies f_1, f_2, and f_3, where each sinusoid is composed of $M = 100000$ samples. **From now on we adopt such an approximation**.

B.1) Write the model of $V[n]$ as a standard w.s.s. harmonic process, composed of 3 **real** sinusoids with same amplitude 2 and frequencies f_1, f_2, and f_3.

At the receiver side (output of the Gaussian channel) we have

$$Y[n] = V[n] + W[n]\,, \quad n = 1, \ldots, 100000\,,$$

where $W[n]$ is a white Gaussian noise, centered, with variance $\sigma_W^2 = 1$ (notice that the variance is now given!).



Given the outcome $y[1], \ldots, y[100000]$, of the noisy harmonic process $Y[n]$:

B.2) Propose a spectral estimation method (based on $y[1], \ldots, y[100000]$) in order to identify the three frequencies f_1, f_2, and f_3 (WITHOUT describing the method in details). Justifying precisely your choice;

B.3) Describe in details the proposed method (based on $y[1], \ldots, y[100000]$). Write every step of your method (as a bullet list) clearly indicating the input and output of each step.

B.4) Indicate if the available number of samples $y[1], \ldots, y[100000]$ are sufficient to apply the proposed method. Justify precisely your answer.

Bonus Questions

B.5) Consider now $V[n]$ to be the very particular type of harmonic process generated by the frequency modulation (and not its approximation to a standard harmonic process), that is, a signal where each 100 samples correspond to a sinusoid with a given frequency (f_1, f_2, or f_3). If you want to apply the method you have proposed, what should you be careful of when identifying the frequencies?

Solution 3.

A) Amplitude Modulation

$$Y[n] = X[n] + W[n]\,, \quad n = 1, \ldots, 1000\,.$$

Call $\boldsymbol{Y} = [Y[1], \ldots, Y[1000]]$, $\boldsymbol{y} = [y[1], \ldots, y[1000]]$, $\boldsymbol{X} = [X[1], \ldots, X[1000]]$, $\boldsymbol{x} = [x[1], \ldots, x[1000]]$, and $\mathcal{A}$ the set of all the possible combination of the values a_1, a_2, a_3 in sequences of length 1000.

A.1) $F_{\boldsymbol{Y}}(\boldsymbol{y}) = P(\boldsymbol{Y} \le \boldsymbol{y})$ can be written as the marginal of the complete cumulative distribution $P(\boldsymbol{Y} \le \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x})$

$$\begin{aligned}
P(\boldsymbol{Y} \le \boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathcal{A}} P(\boldsymbol{Y} \le \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x}) &\overset{\text{Bayes}}{=} \sum_{\boldsymbol{x}\in\mathcal{A}} P(\boldsymbol{Y} \le \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}) \\
&\overset{\text{i.i.d. symbols}}{=} \sum_{\boldsymbol{x}\in\mathcal{A}} P(\boldsymbol{Y} \le \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) \prod_{n=1}^{1000} P(X[n] = x[n]) \\
&\overset{\text{white noise}}{=} \sum_{\boldsymbol{x}\in\mathcal{A}} \prod_{n=1}^{1000} P(Y[n] \le y[n] \mid X[n] = x[n]) \prod_{n=1}^{1000} P(X[n] = x[n]) \\
&= \sum_{\boldsymbol{x}\in\mathcal{A}} \prod_{n=1}^{1000} P(Y[n] \le y[n] \mid X[n] = x[n]) P(X[n] = x[n]) \\
&= \prod_{n=1}^{1000} \sum_{x\in\{a_1,a_2,a_3\}} P(Y[n] \le y[n] \mid X[n] = x) P(X[n] = x)\,,
\end{aligned}$$

where $P(Y[n] \le y[n] \mid X[n] = x)$ is a cumulative distribution of a Gaussian random variable, with mean x and variance σ_W^2, and $P(X[n] = x) \overset{\text{i.i.d.}}{=} P(X = x)$.

A.2) $P(Y[n] \le y[n] \mid X[n] = x)$ is a cumulative distribution of a Gaussian random variable, with mean x and variance σ_W^2, therefore is differentiable and the derivative is a Gaussian density $f(y[n]) = \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(y[n]-x)^2}{2\sigma_W^2}}$.

Consequently

$$\begin{aligned}
f(\boldsymbol{y}) &= \prod_{n=1}^{1000} \sum_{x\in\{a_1,a_2,a_3\}} \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(y[n]-x)^2}{2\sigma_W^2}} P(X = x) \\
&= \prod_{n=1}^{1000} \frac{1}{\sqrt{2\pi\sigma_W^2}} \left(e^{-\frac{(y[n]-a_1)^2}{2\sigma_W^2}} P(X = a_1) + e^{-\frac{(y[n]-a_2)^2}{2\sigma_W^2}} P(X = a_2) + e^{-\frac{(y[n]-a_3)^2}{2\sigma_W^2}} P(X = a_3) \right) \\
&= \prod_{n=1}^{1000} \frac{1}{\sqrt{2\pi\sigma_W^2}} \left(e^{-\frac{(y[n]-a_1)^2}{2\sigma_W^2}} \pi_1 + e^{-\frac{(y[n]-a_2)^2}{2\sigma_W^2}} \pi_2 + e^{-\frac{(y[n]-a_3)^2}{2\sigma_W^2}} \pi_3 \right)\,,
\end{aligned}$$

where $P(X = a_i) = \pi_i$, $i = 1, 2, 3$.

A.3) We have to de-noise a discrete valued signal, therefore we shall apply the classification method: Model the noisy signal as a parametric mixture model; Compute the parameters of the model using the EM algorithm; Estimate the most likely discrete valued signal using a maximum *a priori* approach.

Remark: Applying a Wiener filter is a wrong approach since the resulting de-noised signal will be a smoothed, therefore continuous valued, version of the noisy signal.

A.4) Your are given the received (noisy) signal samples $y[1], \ldots, y[1000]$, and the number of source symbols $K = 3$. As presented above the de-noising is divided into 3 parts:

I) Modeling

The noisy signal $Y[n]$ can be modeled using a parametric probability density, corresponding to a mixture of densities. More precisely, given $N = 1000$ samples of the noisy signal, call $\boldsymbol{Y} = [Y[1], \ldots, Y[1000]]$, and $\boldsymbol{y} = [y[1], \ldots, y[1000]]$. Then, as computed above

$$f(\boldsymbol{y}) = \prod_{n=1}^{1000} \frac{1}{\sqrt{2\pi\sigma_W^2}} \left(e^{-\frac{(y[n]-a_1)^2}{2\sigma_W^2}} \pi_1 + e^{-\frac{(y[n]-a_2)^2}{2\sigma_W^2}} \pi_2 + e^{-\frac{(y[n]-a_3)^2}{2\sigma_W^2}} \pi_3 \right) .$$

The parameters of the model are $\boldsymbol{\theta} = \{a_1, a_2, a_3, \sigma_W^2, \pi_1, \pi_2, \pi_3\}$.

II) Parameter Estimation

Estimation of the parameter is carried out by maximisation of the likelihood function, here given by

$$h(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{n=1}^{1000} \frac{1}{\sqrt{2\pi\sigma_W^2}} \left(e^{-\frac{(y[n]-a_1)^2}{2\sigma_W^2}} \pi_1 + e^{-\frac{(y[n]-a_2)^2}{2\sigma_W^2}} \pi_2 + e^{-\frac{(y[n]-a_3)^2}{2\sigma_W^2}} \pi_3 \right) ,$$

that is, given $\boldsymbol{y}$, $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} h(\boldsymbol{\theta}; \boldsymbol{y})$, under that constraints $\sigma_W^2 > 0$ and $\pi_1 + \pi_2 + \pi_3 = 1$.

Such a maximisation doest not admit en explicit solution and needs to be performed with an iterative algorithm such as the EM algorithm.

III) Estimation of the most likely discrete valued signal

Given the received (noisy) signal samples $y[1], \ldots, y[1000]$, and the estimated parameters $\widehat{\boldsymbol{\theta}}$, the most likely discrete valued signal is estimated by maximization of the *a posteriori* distribution

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) \mathrm{P}\left(\boldsymbol{X} = \boldsymbol{x}\right)}{f(\boldsymbol{y})} ,$$

that is

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{y}) .$$

Here we have (see answers A.1 and A.2)

$$f(\boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) = \prod_{n=1}^{1000} \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-\frac{(y[n]-x[n])^2}{2\sigma_W^2}} , \quad \mathrm{P}\left(\boldsymbol{X} = \boldsymbol{x}\right) = \prod_{n=1}^{1000} \mathrm{P}\left(X = x[n]\right) ,$$

under the constraint that $x[n] \in \{a_1, a_2, a_3\}$.

Given that $f(\boldsymbol{y})$ does not depend on $\boldsymbol{x}$, then the solution is given by

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \prod_{n=1}^{1000} e^{-\frac{(y[n]-x[n])^2}{2\sigma_W^2}} \mathrm{P}\left(X = x[n]\right) .$$

Such a maximization does not admit an explicit solution and needs to be performed using an iterative algorithm such as the Viterbi algorithm.

B) Frequency Modulation

B.1) $V[n]$ as w.s.s., real, harmonic process

$$\begin{aligned} V[n] =& 2\cos(2\pi 0.044n+\theta_1)+2\cos(2\pi 0.0494n+\theta_2)+2\cos(2\pi 0.0523n+\theta_3) \\ =& e^{i2\pi 0.044n+\theta_1}+e^{-i2\pi 0.044n-\theta_1}+e^{i2\pi 0.0494n+\theta_2}+e^{-i2\pi 0.0494n-\theta_2}+e^{i2\pi 0.0523n+\theta_3}+e^{-i2\pi 0.0523n-\theta_3}, \end{aligned}$$

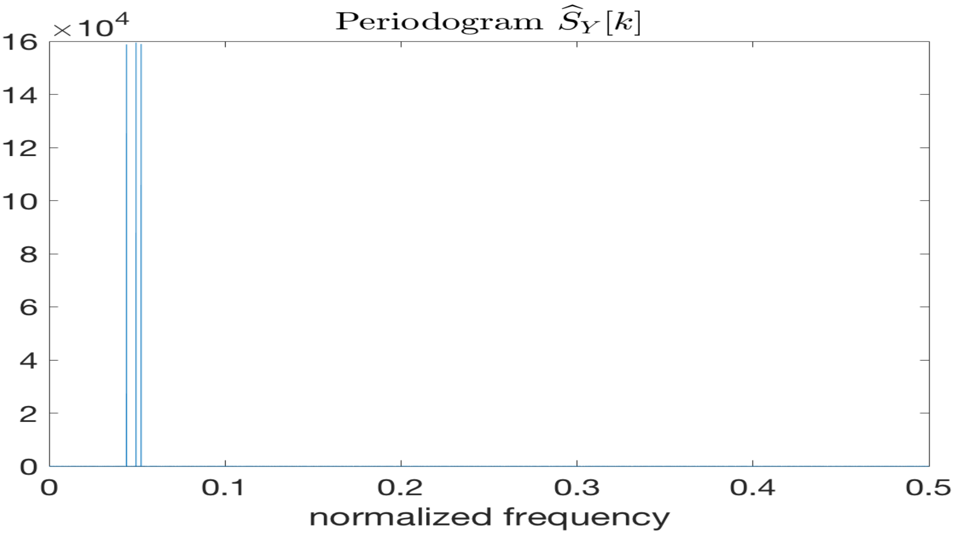
where θ_1, θ_2, and θ_3 are i.i.d. random variable uniformly distributed over $[0, 2\pi]$.

B.2) The signal is noisy with relatively high level of noise. Consequently it is not possible to directly apply the line spectra estimation method seen in class, namely the annihilating filter method. Now we have two possible approaches

- De-noise the signal using the Wiener filter and then apply the annihilating filter method;
- Use another approach less critical in noise, such has the periodogram. Once computed the periodogram the frequencies are detected by thresholding.

The first approach has a higher computational burden since it requires to process the signal twice: first de-noise the signal, then estimate the frequencies. Given the high number of samples, we can expect the Wiener filter de-noising to be efficient. Therefore, the annihilating filter will be applied on a noiseless harmonic signal and will still have the advantage of providing accurate results (with fewer samples one has to be careful that after the de-noising procedure **we do not have a noiseless harmonic signal but a de-noised harmonic signal**, that is, a signal that might differ from a pure harmonic signal!).

The second method, the periodogram, has a lower computational burden that then first method. Given the high number of samples, the bias of the periodogram will be highly reduced and the resolution will be high, providing accurate results. The figure below gives an idea of the periodogram estimate, where the three (six) spectral lines can be clearly distinguished.



Finally, the periodogram is, in such a context, an optimal method.

B.3) I) Compute the periodogram

Given the samples $y[1], \ldots, y[100000]$, the computation of the periodogram is quite straightforward

$$\widehat{S}_y[k] = \frac{1}{N} \left| \sum_{n=1}^{100000} y[n] e^{-i2\pi \frac{(n-1)(k-1)}{N}} \right|^2 , \quad k = 1, \ldots, N\,.$$

II) Perform a thresholding to detect the lines of the spectrum.

We can set the threshold to be .9 the maximum value detect. Given that the number of lines is known (number of possible different symbols), if we do not detect the expected number of lines, we can iteratively lower the threshold.

B.4) The smallest difference between the frequencies is $\Delta = 0.0029$. In order to be able to spectral lines thee number of samples must satisfy

$$N > \frac{1}{0.0029} = 344.83\,.$$

Here we have $N = 100000$, a number of samples satisfying the above constraint.

B.4) Bonus question.

The very particular harmonic process can be seen as a standard harmonic process where each sinusoids is windowed by square windows: The window of a sinusoid with frequency f_k is equal to 1 for 100 samples when the symbol takes the value s_k, $k = 1, 2, 3$.

Such a windowing will highly affect the resolution of the periodogram and therefore the minimum number of samples required.

Exercise 4. A HOBBIST TRANSMISSION SYSTEM (20 PT)

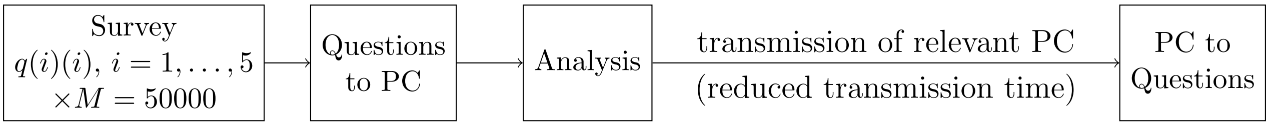
A market survey specialist happens to be a transmission system hobbist. He/She decides to design a system to transmit the results of his latest survey composed of 5 questions that we shall call $q(1)$, $q(2)$, $q(5)$, $q(4)$, and $q(5)$. The answers to the questions corresponds to numerical values.

Given that the survey sample size is of 50000 individuals (independent and of the same social group), he/she quickly realizes that the transmission time required to transmit the survey results is way too long!

In total frustration, He/She decides to contact you looking for some help.

Having attended the statistical signal processing class, you suggest him to use principal component analysis in order to reduce the complexity of the transmission. Unfortunately, He/She has no idea about principal component analysis. Consequently you have to do the job!

The idea is to have a system like the following:


Survey
$q(i)(i)$, $i = 1, \ldots, 5$
$\times M = 50000$
Questions to PC
Analysis
transmission of relevant PC
(reduced transmission time)
PC to Questions

Show in detail how to apply the principal component analysis to this problem, and more precisely:

1) Check if the necessary conditions to apply the PCA method are satisfied or eventually make appropriate assumptions.

2) Write the equations describing the method, **clearly indicating the dimensions of the matrices**.

3) Precisely describe the steps necessary to implement the method, **referring to the above block diagram** and according to the available data($q(i)$, $i = 1, \ldots, 5$, $\times 50000$). Write every step of your method (as a bullet list) clearly indicating the input and output of each step (if we plug in such steps into a computational software we must then obtain the desired result .. so don't miss any step! In particular the input, the executed operation with corresponding equations, and the output of each step has to be clear).

Remark: You are not asked to describe the PCA, but rather to describe how the PCA can be applied to solve this very particular problem!

Solution 4.

The sample size of the survey is $M = 50000$ (number of individuals that have been asked the questions), the number of questions is $N = 5$.

1) As specified, the individuals are supposed to be mutually independent and of the same social group. Consequently, the $M = 50000$ sets can be modeled as an i.i.d. process. Such a process is *a priori* not zero mean but it can be easily centered. Therefore, the zero mean - w.s.s. condition for the application of the PCA is satisfied.

2+3) We have:

Questions to PC

- Set $\boldsymbol{q}_k = [q_k(1), \dots, q_k(5)]$, $k = 1, \dots, 50000$;
- Center the variables: $\bar{\boldsymbol{q}}_k = \boldsymbol{q}_k - \boldsymbol{m}_q$ (dimensions 1×5), $k = 1, \dots, 50000$, where $\boldsymbol{m}_q = \dfrac{1}{50000} \displaystyle\sum_{n=1}^{50000} \boldsymbol{q}_n$ (dimensions 1×5);
- Estimate the correlation matrix $\widehat{\boldsymbol{R}}_Q = \dfrac{1}{50000} \displaystyle\sum_{n=1}^{50000} \boldsymbol{q}_n^t * \boldsymbol{q}_n$ (dimensions 5×5);
- Diagonalize the correlation Matrix $\boldsymbol{V}^t \widehat{\boldsymbol{R}}_Q \boldsymbol{V} = \boldsymbol{\Lambda}$ (dimensions 5×5);
- Compute the principal components $\boldsymbol{z}_k = \boldsymbol{q}_k \boldsymbol{V}$ (dimensions 1×5), $k = 1, \dots, 50000$.

Analysis

- Determine which are the most relevant eigenvalues by looking at $\boldsymbol{\Lambda}$. We suppose in the following that there are 3 of them;
- Call $\widetilde{\boldsymbol{z}}_k = [z(1)_k, z(2)_m, z(3)_m]$ (dimensions 1×3), $k = 1, \dots, 50000$, the vectors of the most relevant principal components.

Transmission

- Transmit $\widetilde{\boldsymbol{z}}_k = [z(1)_k, z(2)_m, z(3)_m]$, $k = 1, \dots, 50000$ (reducing the transmission by a factor $2/5$);
- Transmit the mean $\boldsymbol{m}_q$ (**this is a fundamental step, otherwise it is not possible to reconstruct the questions**);
- Transmit the eigenvector matrix $\boldsymbol{V}$ (**this is a fundamental step, otherwise it is not possible to reconstruct the questions**).

PC to questions

- Obtain the centered questions from the principal component

$$\widehat{\bar{\boldsymbol{q}}}_k = [\widehat{q}_k(1), \dots, \widehat{q}_k(5)] = [z(1)_k, z(2)_k, z(3)_k, 0, 0]\boldsymbol{V}^t\,, \quad k = 1, \dots, 50000\,.$$

- Add the mean to obtain the questions $\widehat{\boldsymbol{q}}_k = \widehat{\bar{\boldsymbol{q}}}_k + \boldsymbol{m}_q$, $k = 1, \dots, 50000$.

Grade Scale.

The exams accounts for a total of 65 points (exact response to each question) + 3 bonus points.

The grading has been done on a 59 points scale (59 points = 6/6), according to the following formula

$$\text{grade over } 6 = 1 + (5 * \text{points}/59)$$

and then rounded to .5 steps, that is

$$\text{rounded grade over } 6 = (\text{round-to-0-digit}(2 * \text{grade over } 6))/2$$

The result is then constraint to be at maximum 6.