ECEN321: Engineering Statistics
Laboratory Session 3
Estimation

Due: 9:00 a.m., Monday 25 May 2020

Write a report according the format outlined in the first lab and describe the following lab.

You often have some idea of the family of probability distribution to which a random variable belongs. You may believe the outcome of your experiment has a Gaussian distribution, or a Laplacian distribution, for example. However, you usually do not know the *parameters* of the distribution, and so you may need to estimate them. In this lab we first will try to get some intuition about the behavior of random variables, and then try to estimate parameters of a distribution of a real-world signal.

1 Gaussian Random Variables

Let us consider a Gaussian random variable with a mean of 10 and a variance of 4. We first generate the data and, from the data, will try to recover the mean and the variance. We consider two estimation methods, one with 10 experiments and one with 10 000 experiments (so we just generate 10 and 10 000 samples, drawn from the same Gaussian distribution). So the first estimation method involves 10 *realisations* of the random variable and the second 10 000 realisations of the random variable. To see how good the obtained estimates are, we run each method 1000 times (you could do this all at once by generating a 10×1000 matrix and a $10,000 \times 1000$ matrix).

1. For both methods compute the 1000 sample means.
2. For both methods compute the 1000 sample variances.
3. Plot four simple scatter plots with the horizontal axis showing the estimation index (from 1 to 1000).
4. Plot the corresponding four histograms.
5. Explain the shape of the distribution for the variances qualitatively.
6. Explain the differences between the method with 10 experiments and the method with 10 000 experiments based on the theory discussed in lectures. Discuss all aspects that you can explain. Your analysis should be based on mathematical logic, with equations, and not just a *qualitative* story.

2 The Probability Distribution of Speech

In this task we will study the distribution of the samples of the speech signal. In this type of context, the distribution is called a *model* of the signal. If you want to generate synthetic speech or if you want to encode the speech signal (like your mobile phone does), it is beneficial to know the distribution (to have a model) of the speech samples. Because it simplifies mathematical derivations, it is often assumed that the distribution of the samples is Gaussian. We will check here if that is reasonable.

For our hypotheses we consider distributions from the family of generalized normal distributions, which are discussed in more detail at `http://en.wikipedia.org/wiki/Generalized_normal_distribution` The generalized normal distribution is described by:

$$f(x;\alpha,\beta,\mu) = \frac{\beta}{2\alpha\Gamma(1/\beta)}\, e^{-(|x-\mu|/\alpha)^{\beta}}, \tag{1}$$

where x is the variable α, β, and μ are parameters, and where $\Gamma(\cdot)$ is the gamma function. A particular model is now specified by the parameters α , β and μ. For $\beta = 2$ the distribution is precisely the Normal distribution (Gaussian) and for $\beta = 1$ it is a Laplacian distribution. We are particularly interested in the value of β.

1. The file `speech.pcm` contains a speech signal segment in 16 bit integer format; the signal is sampled at 16 kHz (so-called "wide-band speech"). Extract the data and make a plot of them to verify that you did the extraction correctly.

2. Measure the variance σ^2 of the samples of the speech signal.

3. Argue that assuming that the samples are independent is ok for the purpose of estimating β and α.

4. Assume throughout that $\mu = 0$.

5. Approximate $\alpha = \sqrt{2\sigma^2}$ and find the ML estimate of β. You do that by a trying a set of values for β and selecting the value for β for which the probability density of the observed data is highest (remember the joint density of all observations is the same as the multiplication of the densities of the individual observations if we assume they are independent). Note that the probability densities are likely to be small and you will run out of machine precision if you use them directly. Instead, maximize the log of the probability densities (note that multiplication then becomes an addition).

6. Given the β you found, find the ML estimate of α.

7. Keep going back and forth between estimating α and β until it converges (you want to write a loop). If it does not converge use a grid search to find the ML estimate for α and β. What are the final values of the parameter pair $\theta = [\alpha, \beta]$?