Characterization of a population parameters

Davide Pedranz, Mat. number 189295
davide.pedranz@studenti.unitn.it

***Abstract*—An electronic probe measures the output of a circuit which is known to generate independent identically distributed samples taken from a logistic distribution. The probe is affected by a Gaussian white noise and a sinusoidal bias. Given the measures data set, we want to estimate the mean and its confidence interval in order to decide whether to keep or discard the circuit.**

I. Introduction

Circuitry are widely used in every electronic device. In order for those devices to work properly, high quality circuitry is needed. The quality of a circuitry can be evaluated measuring some of its properties. Unfortunately, each measurement is subject to some random noise or bias. In order to achieve reliable results, the stochastic nature of the measuring process must be taken into account.

In this case, we have a circuitry which is known to generate independent identically distributed (i.i.d.) samples taken from a population with a logistic distribution. The electronic probe used for the measures is known to introduce a white Gaussian noise and a sinusoidal bias. The actual measures can thus be described as:

$$\begin{aligned} x_t &= Asin(2\pi ft) + Y + Z, \\ Y &\sim Logis(\mu, s),\ Z \sim N(0, \sigma_z). \end{aligned} \tag{1}$$

Our aim is to estimate the five parameters of the model given the measures data set. In particular, we are interested in the mean μ of the logistic distribution Y. In addition, a confidence interval for the $\hat{\mu}$ is required.

The given data set contains $N = 60\,000$ points from a sampling done at $F = 20\,\text{kHz}$ for $3\,\text{s}$. We assume the frequency f of the sinusoid to be an integer divisor of the sampling frequency and to be low, in the order of ten/hundreds of Hz.

We will estimate one parameter at a time, use it to clean the samples, then focus on the next one until the last one. In particular, we will start from μ, then estimate f and A, and finally s and σ. As the last step, we will estimate the confidence interval for $\hat{\mu}$.

II. First parameter - Mean μ

We start from the estimation of the mean μ of the logistic distribution Y. We observe that:

- the Gaussian Z has zero mean, thus $E[Z] = 0$;
- the data set contains an integer number of periods of the sinusoid, which implies that its mean over the entire data set is zero $E[Asin(2\pi ft)] = 0$.

Thus, we can build an unbiased estimator $\hat{\mu}$ by taking the average on the whole data set:

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x_t = 3.00 \tag{2}$$

It is easy to prove that $\hat{\mu}$ in Equation (2) is unbiased:

$$\begin{aligned} E[\hat{\mu}] &= E\Big[\frac{1}{n}\sum_{t=1}^{n}\big(Asin(2\pi ft) + Y + Z\big)\Big] \\ &= \frac{1}{n}\sum_{t=1}^{n} E[Asin(2\pi ft)] + \frac{1}{n}nE[Y] + \frac{1}{n}nE[Z] \\ &= 0 + \mu + 0 = \mu. \end{aligned}$$

III. Second Parameter - Frequency f

The second parameter in our estimation is the frequency f of the sinusoid. First, we clean the data set by subtracting the estimation of the mean μ. The new data set has zero mean and can be described in a similar way to the original one in Equation (1):

$$\begin{aligned} x_t &= Asin(2\pi ft) + Y + Z, \\ Y &\sim Logis(0, s),\ Z \sim N(0, \sigma_z). \end{aligned}$$

Since Y and Z are independent on each other and produce i.i.d. samples, the correlation of the single random variables is caused only by the sinusoidal noise. S0, we can use the autocorrelation function to estimate the frequency f.

The result of the autocorrelation on a sine function is a cosine with the same frequency, as shown by Equation (3).

$$\begin{aligned} R(\tau) &= E[Asin(2\pi ft) \cdot Asin(2\pi f(t+\tau))] \\ &= \frac{1}{T_c}\int_0^{T_c} A^2(sin(2\pi ft) \cdot sin(2\pi f(t+\tau))\,dt \\ &= \frac{1}{T_c}\Big[A^2\frac{1}{2}cos(2\pi ft)\ t + \frac{1}{8\pi f}sin(2\pi f(\tau+2t)) + \\ &\quad - \frac{1}{8\pi f}sin(2\pi f\tau)\Big]_{t=0}^{t=T_c} \\ &= \frac{1}{2}A^2 cos(2\pi f\tau) \end{aligned} \tag{3}$$

Note that T_c is much greater than $\frac{1}{f}$, so it holds:

$$sin(2\pi f\tau) = sin(2\pi f(\tau + 2T_c)).$$

Using Equation (3), we can compute the autocorrelation function on the data set:

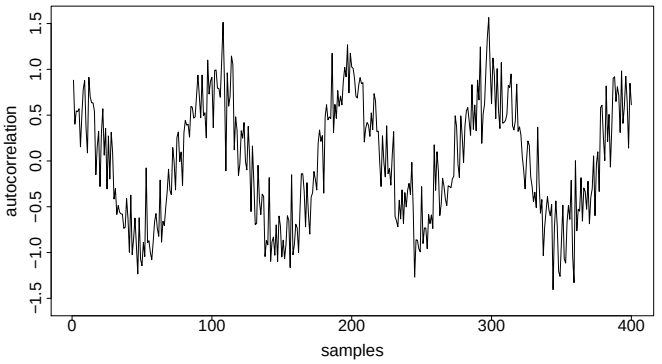

Figure 1. Autocorrelation function of the first 400 points of the sample. It is easy to notice the underlying a cosine function.

$$\begin{aligned} R(\tau) &= E[X(t) \cdot X(t+\tau)] \\ &= \frac{1}{2}A^2 cos(2\pi f\tau) + \mathrm{Var}[Y] + \mathrm{Var}[Z]. \end{aligned} \tag{4}$$

Figure 1 shows the autocorrelation function of the first 400 points of the data set. As expected, the function has the shape of a cosine function. The residual noise due to the variances of Y and Z.

It is easy to see from the graph that each period of the sinusoid contains about 100 samples. To get the precise number, we use a numerical approach. We denote with lag half of the period of the sinusoid. For an interval of our rough estimation of the lag (in this case 50), we compute the average of the absolute value of the peaks (positive and negative) as a function of the lag's guess. Then, we take the value that maximizes this function as the estimated lag.

$$lag = \underset{l}{\mathrm{argmax}} \sum_{\tau=1}^{\left\lfloor \frac{N}{2 \cdot l} \right\rfloor} |R(\tau + l)| + |R(\tau + 2 \cdot l)|$$

Now the frequency f can be estimated as:

$$\hat{f} = \frac{F}{2 \cdot lag} = \frac{20\,\mathrm{kHz}}{2 \cdot 50} = 200\,\mathrm{Hz}$$

IV. Third Parameter - Amplitude A

The amplitude A of the sinusoid can be computed in a similar way. Equation (3) describe the relationship between the original sine signal and its autocorrelation function. To estimate the amplitude of the autocorrelation function A_{cos}, we take the average of the absolute value of the theoretical highest and lowest points of each period of the cosine. Formally:

$$\begin{aligned} n &= \left\lfloor \frac{2 \cdot lag}{N} \right\rfloor, \\ A_{cos} &= \frac{1}{n} \sum_{\tau=1}^{n} |R(\tau + lag)| + |R(\tau + 2 \cdot lag)| = 0.81 \end{aligned} \tag{5}$$

As for the case of the frequency, we take the mean over all peaks to reduce the noise due to the variances of Y and Z.

Then, we can easily derive the estimation of the original amplitude A:

$$\hat{A} = \sqrt{2 \cdot A_{cos}} = 1.27.$$

V. Last Parameters - s and σ_n

We remove the sinusoid from the data set by subtracting the theoretical sine function with frequency equal to the estimated frequency f and amplitude equal to the estimated amplitude A. Now the samples can be described as:

$$\begin{aligned} X &= Y + Z, \\ Y &\sim Logis(0, s),\ Z \sim N(0, \sigma_z). \end{aligned}$$

To estimate the remaining parameters s and σ, we can analyze the moments of $X = Y + Z$.

Both Y and Z have zero mean, so the first moment does not provide any useful information about the single distributions.

Since Y and Z are independent, the variance of the sum is equal to the sum of the variances, formally:

$$\mathrm{Var}[Y + Z] = \mathrm{Var}[Y] + \mathrm{Var}[Z] \tag{6}$$

Both a logistic and a Gaussian distribution are symmetric with respect to the origin, so the skewness of X, Y and Z is zero.

The excess kurtosis of a Logistic distribution is always 1.2, while the Gaussian distribution one is always 0:

$$\mathrm{Kurt}[Y] - 3 = 1.2,\ \mathrm{Kurt}[Z] - 3 = 0.$$

We also know that the excess kurtosis of the sum of independent random variables for which the fourth moment exists has the following property, according to [1]:

$$\mathrm{Kurt}[Y + Z] - 3 = \frac{\sigma_y^4(\mathrm{Kurt}[Y] - 3) + \sigma_z^4(\mathrm{Kurt}[Z] - 3)}{(\sigma_y^2 + \sigma_z^2)^2},$$

where σ_y and σ_z denote respectively the standard deviation of Y and Z. We can thus compute the excess kurtosis of X:

$$\mathrm{Kurt}[X] - 3 = \frac{1.2 \cdot \sigma_y^4}{(\sigma_y^2 + \sigma_z^2)^2}. \tag{7}$$

From Equation (6) and Equation (7) we obtain a system of two equations in two variables. $\mathrm{Var}[X]$ and $\mathrm{Kurt}[X] - 3$ are computed on the data set using the standard well known estimators. The system has two solutions, but we are interested only in the positive one.

$$\begin{cases} \hat{\sigma}_y^2 = \sqrt{\frac{(\mathrm{Kurt}[X]-3) \cdot \mathrm{Var}[X]^2}{1.2}} \\ \hat{\sigma}_z^2 = \mathrm{Var}[X] - \hat{\sigma}_y^2 \end{cases}$$

The estimate of s and σ_z becomes now trivial:

$$\hat{s} = \frac{\hat{\sigma}_y}{\pi}\sqrt{3} = 4.49,\ \ \hat{\sigma}_z = 0.97.$$

Table I
SUMMARY OF THE ESTIMATED PARAMETERS

parameter		value
μ	mean of the logistic Y	3.00
f	frequency of the sinusoid	200 Hz
A	amplitude of the sinusoid	1.27
s	scale of the logistic	0.49
σ	s.d. of the Gaussian	0.96

VI. MEAN CONFIDENCE INTERVAL

The data set contains a sinusoidal noise, which makes the samples sample not identically distributed. In order to compute a small confidence interval with a high probability, we need to transform the data set to obtain i.i.d. samples. The used technique is the batch means.

We observe that the mean operator is linear. This allows to compute the mean of the original data set by dividing the samples in batches of some fixed size, computing the mean of each batch and finally the mean of the means.

If we take the size of the batch as a multiple of the wavelength of the sinusoid and compute the batches, we obtain a population of random variables i.i.d., since the sum of any sinusoidal signal over a single period is zero. Since the size is still high (300 elements), we can treat the new population as normally distributed. We computed two confidence intervals for a confidence of 95% and 99%:

$$\begin{aligned} P[2.94 < \mu < 3.06] &= 0.95 \\ P[2.92 < \mu < 3.08] &= 0.99 \end{aligned} \tag{8}$$

A second possibility is to consider the data set without the sinusoidal noise. Removing the sinusoidal signal is subject to some noise which increases the variance of the new data set. On the other, we can directly treat the population as normally distributed and use the confidence intervals for a Gaussian population. The result of the computation gives practically the same numbers as the batch means technique (the difference is less 10^{-10}). The high number of samples probably compensate the extra noise introduces by the sinusoid's removal.

VII. CONCLUSION

In this work, we solved the problem of estimating the parameters of a given model using a sample of measures. This type of analysis should be used for every measure that requires some reliability, since every probe is subject to somer random noise.

The estimated values for each parameter are summarized in Table I. The confidence intervals defined in Equation (8) are reported below for completeness:

$$\begin{aligned} P[2.94 < \mu < 3.06] &= 0.95 \\ P[2.92 < \mu < 3.08] &= 0.99 \end{aligned}$$

First, we estimated the value for the mean parameters, which was not affected by the others. Secondly, we used the auto correlation function to isolate and remove the sinusoidal noise. Thirdly, we exploited the different shapes and moments of the Gaussian and the logistic distribution to separate them from each other. Finally, we computed a confidence interval for the mean using the technique of the batch means.

REFERENCES

[1] Wikipedia, "Kurtosis — Wikipedia, The Free Encyclopedia," 2016, online; accessed 5-November-2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Kurtosis&oldid=747942450