M1 - Marine Physics **2019–2020**

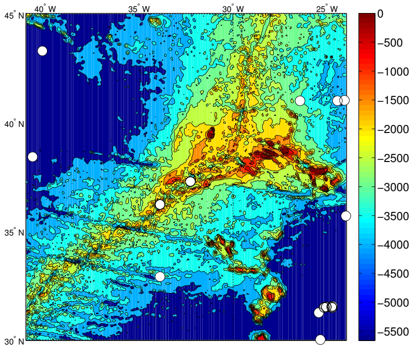
Data Analysis

#1 Statistical Methods

1 Probability Density Function (pdf)

You will use real data from a bottom current meter on the Mid-Atlantic Ridge. The file contains variables **time**, **u** and **v** and is available here :

`http://stockage.univ-brest.fr/~gula/TS1/current.mat`



1. Load the variables from the file
   matlab : `load current.mat`, python : `scipy.io.loadmat('current.mat')`
2. Plot the time series of **u** and **v**
3. Plot the histogram of **u** on the interval $[-0.5, 0.5]$ using a bin width of 0.001
4. Compare the results using different bin widths (0.001, 0.01, 0.02, 0.05, 0.1). Is there an optimal choice ?
5. The optimal number of bins k is often estimated as $k = 1 + \log_2 N$, where N is the number of samples. How does it compare to your optimal choice ?
6. Find a way to normalize the histogram to plot the probability density function (pdf). *i.e.* all the histograms should collapse on one curve, independently of the number of elements N and of the bin widths you choose.
7. Write a function `mypdf` returning the pdf of `u`. The function will have on input the bins vector and `u`.
8. Compute the cdf $C(u)$ [Use `cumsum`].
9. Compute the interval $[-a, a]$ on which we have 68% chance of finding x. Same question for 95%, and 99%.

2 Statistics

1. Compute the mean of `u`
2. Compute the median of `u`
3. Compute the standard deviation of `u` [use `std`]. Standard deviation of a variable x_k can be computed using the unbiased estimation

$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n-1} \tag{1}$$

or the biased one

$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n} \tag{2}$$

The denominator is $n-1$ in the unbiased case because there is only $n-1$ degrees of freedom to estimate σ since one is used to estimate $\bar{x}$.

4. Compute the skewness of `u` using i) the matlab/python function *(belonging to the signal toolbox in matlab or the scipy.stats module in python)* ii) directly from its definition

$$\texttt{skewness} = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^3}{(n-2)\,\sigma^3} \tag{3}$$

Skewness is a measure of asymmetry of the pdf. Symmetric pdfs have zero skewness. The factor $n-2$ (instead of n) ensures an unbiased estimator. Two degrees of freedom are already used to estimate the skewness : one for $\bar{x}$, the other for σ.

5. compute the kurtosis of `x`

$$\texttt{kurtosis} = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^4}{(n-2)\,\sigma^4} \tag{4}$$

Kurtosis is a measure of how extreme events are important. Kurtosis for gaussian is 3.

3 Normal distribution

1. Generate a random variable `x`, normally distributed, with the same number of elements N, the same mean and the same standard deviation than **u**.

   matlab : `randn(N)`, python : `np.random.randn(N)`

2. Plot its pdf
3. Compare it to the analytical pdf of a normal distribution
4. Compare it to the pdf of u
5. Compare the first four moments (mean, std, skewness, kurtosis) of x and compare them with the moments of u
6. Redo the same procedure (plot the pdf and compute the moments) for a random variable $\texttt{x}_{50}$ following the same distribution but with a number of elements N = 50. What do you see ?
7. compute the cdf $C(x)$
8. Check if x and u are Gaussians using the following statistical tests : the Shapiro-Wilk test, the D'Agostino's Test, and the Anderson-Darling est.