Data Analytics for Coastal Systems

Georgios Boumis, Ph.D.
Assistant Professor
Department of Civil and Environmental Engineering
The University of Maine

Abstract. This course introduces statistical methods for analyzing environmental data, where students develop R programming skills through hands-on work on processing real-world data from Maine's coastal environments. Emphasis is given in datasets such as sea level observations, temperature and sea-surface pressure measurements, as well as seasonal wind speed records. The course advances systematically from basic data visualization and descriptive statistics to predictive modeling, probability distribution fitting via maximum likelihood estimation, and parametric uncertainty quantification. It culminates with the application of extreme-value theory and trend detection for evaluating rare coastal events, particularly within the context of non-stationarity due climate change-driven sea-level rise.

Contents

5 Parameter Estimation and Goodness-of-fit

Recall from **Chapter 4** that for the daily average sea-surface pressure measurements at Cutler Farris Wharf, ME, in 2015 (Figure 1), we failed to reject the null hypothesis that the data "come" from a normal distribution (see Shapiro-Wilk test). Therefore, a normal probability density function (pdf) is a suitable function to model the probability of these data. However, we do not want to use any normal pdf, but rather the specific one that arises from the combination of parameters μ and σ that maximizes the fit (Figure 2). The process of finding the values of the parameters that achieve the maximal fit is called "fitting a distribution to the data". This process involves two key ingredients: 1) a measure of "fit", and 2) a mathematical procedure to obtain the parameter values.

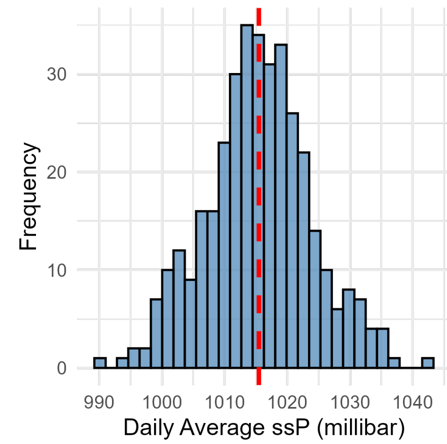

Figure 1: Histogram of daily average sea-surface pressure (ssP) measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line. According to the Shapiro-Wilk test, the data originate from a normal distribution.

5.1 Likelihood ($\mathcal{L}$) Function

A measure of fit for our purpose of fitting a distribution is the so-called **likelihood** ($\mathcal{L}$), which is a function that measures how plausible the parameters of our distribution are given the measurements we have collected. Thus, if we denote the parameters of our distribution by θ and the $i.i.d.$ measurements $(x_1, x_2, \ldots, x_n)$ by x, then the likelihood can also be denoted as $\mathcal{L}(\theta|x)$. Another interpretation of the likelihood is that it measures how probable our data are given different parameter values. Hence, for continuous random variables it follows that:

$$\mathcal{L}(\theta|x) = f(x|\theta) \tag{1}$$

$$\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta), \tag{2}$$

where f is the probability density function with parameters θ. Because each measurement of our sample is independent of each other ($i.i.d.$), Equation 2 can be rewritten as:

$$\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta), \tag{3}$$

where $\prod$ stands for product. This follows from the fact that if two events A and B are independent, then the probability of event $A \cap B$ ("A and B") is given by:

$$P(A \cap B) = P(A) \times P(B). \tag{4}$$

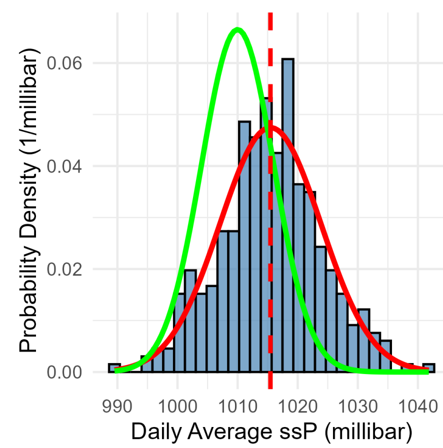

Figure 2: Probability density function (pdf) of normal distribution (red solid line) best fit to daily average sea-surface pressure (ssP) measurements at Cutler Farris Wharf, ME, in year 2015. The mean of the data is shown with a red dashed line. The green solid line denotes another normal pdf with μ and σ different from those of the best fit.

In summary, the likelihood is the product of the pdf evaluated at each measurement and serves as a measure of fit. It is this quantity we want to maximize in order to obtain the optimal set of parameters θ.

5.2 Maximum Likelihood Estimation (MLE)

A mathematical procedure for obtaining the optimal set of parameters is the maximization of the product in Equation 3, i.e., **the maximization of the likelihood**. We will denote the optimal set of parameters obtained by such a procedure as $\hat{\theta}_{MLE}$, or simply $\hat{\theta}$. Because probability densities are small numbers, their product, especially over multiple measurements, becomes such a tiny number that can cause numerical instabilities in the maximization procedure. Therefore, for maximization, it is better to work with the so-called

log-likelihood function (l) instead:

$$l(\theta|x_1, x_2, \dots, x_n) = \log[\mathcal{L}(\theta|x_1, x_2, \dots, x_n)] \quad (5)$$

$$l(\theta|x_1, x_2, \dots, x_n) = \log[\prod_{i=1}^{n} f(x_i|\theta)] \quad (6)$$

$$l(\theta|x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} \log[f(x_i|\theta)]. \quad (7)$$

This is made possible knowing that the logarithmic function is a monotonic (strictly increasing) function and therefore:

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ \mathcal{L}(\theta|x_1, x_2, \dots, x_n) = \underset{\theta}{\text{argmax}}\ l(\theta|x_1, x_2, \dots, x_n). \quad (8)$$

5.3 Short Note: Maximization

Let us briefly refresh our memory on maximization. For this, let us consider the function $f(x) = -(x-3)^2 + 10$, which we will try to maximize. We begin by expanding the expression:

$$f(x) = -(x-3)^2 + 10 \quad (9)$$

$$f(x) = -(x^2 - 6x + 9) + 10 \quad (10)$$

$$f(x) = -x^2 + 6x + 1. \quad (11)$$

Then, to find the x that maximizes $f(x)$, we compute the first the derivative, set it to zero, and solve for x:

$$f'(x) = -2x + 6 \quad (12)$$

$$f'(\hat{x}) = -2\hat{x} + 6 = 0 \Rightarrow \hat{x} = 3. \quad (13)$$

We can verify that this is indeed a maximization point by finding the sign of the second derivative at $\hat{x}$:

$$f''(x) = -2 \quad (14)$$

$$f''(\hat{x}) = -2 \Rightarrow \hat{x} = \text{maximum}. \quad (15)$$

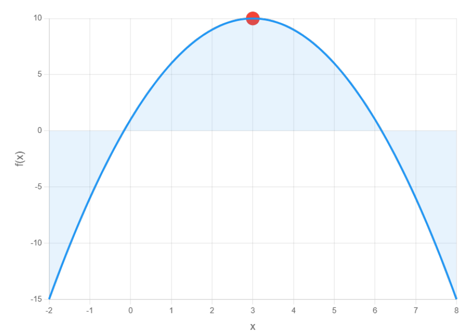

Figure 3: Graph of $f(x) = -(x-3)^2 + 10$ with the red dot showing the maximization point at $(3, 10)$.

5.4 Analytical Solutions of MLE for the Normal Distribution

To fit a normal distribution and find the optimal set of parameters $\theta = (\mu, \sigma)$, we will make use of maximum likelihood estimation as described above. We begin with the likelihood of the normal distribution:

$$\mathcal{L}(\mu, \sigma|x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \quad (16)$$

$$\mathcal{L}(\mu, \sigma|x_1, x_2, \dots, x_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \quad (17)$$

$$(18)$$

Hence, taking the logarithm and finding the log-likelihood leads to:

$$l(\mu, \sigma|x_1, x_2, \dots, x_n) = \log[\frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}] \quad (19)$$

$$l(\mu, \sigma|x_1, x_2, \dots, x_n) = \log[\sigma^{-n}(2\pi)^{-n/2}] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \quad (20)$$

$$l(\mu, \sigma|x_1, x_2, \dots, x_n) = -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. \quad (21)$$

To find the maximum of the log-likelihood with respect to the μ parameter, we set the respective partial derivative to zero and we solve for μ:

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \tag{22}$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \Rightarrow \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{23}$$

Similarly, to find the maximum of the log-likelihood with respect to the σ parameter, we set the respective partial derivative to zero and we solve for σ:

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{24}$$

$$-\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \Rightarrow \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}. \tag{25}$$

5.5 Analytical Solution of MLE for the Rayleigh Distribution

The likelihood of the Rayleigh distribution with scale parameter σ is given by:

$$\mathcal{L}(\sigma|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n}\frac{x_i}{\sigma^2}e^{-\frac{x_i^2}{2\sigma^2}}, \tag{26}$$

and therefore the log-likelihood is:

$$l(\sigma|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n}\log[\frac{x_i}{\sigma^2}e^{-\frac{x_i^2}{2\sigma^2}}] \tag{27}$$

$$l(\sigma|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n}[\log(x_i) - \log(\sigma^2) - \frac{x_i^2}{2\sigma^2}] \tag{28}$$

$$l(\sigma|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n}[\log(x_i)] - 2n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2. \tag{29}$$

To find the maximum of the log-likelihood with respect to the σ parameter, we set the ordinary derivative to zero and we solve for σ:

$$\frac{\partial l}{\partial \sigma} = -\frac{2n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}x_i^2 \tag{30}$$

$$-\frac{2n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}x_i^2 = 0 \Rightarrow -2n\hat{\sigma}^2 + \sum_{i=1}^{n}x_i^2 = 0 \Rightarrow \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}x_i^2}{2n}}. \tag{31}$$

5.6 Numerical Solutions of MLE

Not all MLE problems have closed-form analytical solutions. Although simple distributions such as the normal and Rayleigh have elegant formulas for their parameter estimates, many distributions require **numerical optimization methods**. For example, when trying to maximize the likelihood of the Gamma distribution with respect to the shape parameter (α), the equation involves the digamma function and cannot be solved algebraically. In such cases, we have to rely on iterative numerical algorithms such as, e.g., the Newton-Raphson method. These methods start with an initial guess and iteratively refine the parameter estimates until convergence. Built-in and external functions in R can help us do numerical optimization!