Mathematics for Bioinformatics & Systems Biology

Exercise: Parameter estimation

PROF. DR. JENS TIMMER

Task 1
For the linear regression model

$$y_i = ax_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

the maximum likelihood estimator $\hat{a}$ for a based on N data points (y_i, x_i) is given by

$$\hat{a} = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} x_i^2}. \tag{2}$$

- Implement the estimator.
- Choose $a = 1$, $x_i \sim \mathcal{N}(0, 100)$, and $\varepsilon_i \sim \mathcal{N}(0, 1)$. Generate $M = 1000$ independent estimates $\hat{a}$ for $N = 2, 5, 10, 100, 1000$.
- Does the estimator follow a Gaussian distribution for the different N?

  Hint: Calculate the histogram and the cumulative distribution.
- Calculate the variance of the estimator for $N = 2, 5, 10, 100, 1000$. Display the dependence of the variance and of the standard deviation of the estimator on N graphically.

  Hint: Think about a logarithmic scale.

Task 2
Choose uniformly distributed errors

$$\eta_i \sim U(-b, b)$$

for the model, Eq. (1), and repeat the analysis.

- Choose b such that the variance of η equals 1.

  Hint: The mean of η is zero. $U(-b, b) = 1/(2b)$ between $-b$ and b, zero otherwise. Therefore, $Var(\eta) = \int_{-b}^{b} 1/(2b)\, x^2 dx$. Solve the integral and use V$ar(\eta) = 1$.
- Use the estimator Eq. (2). Understand why this is not the maximum likelihood estimator.

Task 3
Choose Cauchy distributed errors ϕ_i

$$prob_{Cauchy}(\phi_i, \mu, \gamma) = \frac{1}{\pi} \frac{\gamma^2}{(\phi_i - \mu)^2 + \gamma^2}$$

for the model, Eq. (1), and repeat the analysis.

- Choose $\mu = 0$ and $\gamma = 1$.

  Hint: Cauchy distributed random variables for $prob_{Cauchy}(\phi_i, 0, 1)$ can be generated by dividing two independent Gaussian $\mathcal{N}(0, 1)$ random variables.

- Use the estimator Eq. (2). Understand why this is not the maximum likelihood estimator.

Task 4
Compare and understand the results of Tasks 1-3.
Hint: The Cauchy distribution has no finite moments.