Exercise Sheet 1: MLE Simulation Studies

Student-t and Binomial Distributions

Dr. Florian Herzog, FHGR

CDS120 - Uncertainty Quantification

Instructions

This exercise sheet applies the statistical framework and clean notation system introduced in the lecture to concrete simulation studies. You will implement Maximum Likelihood Estimation (MLE) for different distributions and analyze the uncertainty quantification properties through simulation.

Learning Objectives:

- Apply the three-layer framework: True $\rightarrow$ Observed $\rightarrow$ Estimated
- Implement MLE estimators and analyze their distributional properties
- Understand how distributional assumptions affect parameter uncertainty
- Practice uncertainty quantification through simulation studies

Notation Reminder:

- **True (unknown)**: No symbol (e.g., $\boldsymbol{\theta}$, $f(X|\boldsymbol{\theta})$)
- **Observed data**: Tilde ($\tilde{}$) (e.g., $\tilde{X}_i$, $\tilde{y}_i$)
- **Candidate model**: Asterisk ($*$) (e.g., $\boldsymbol{\theta}^*$, $f^*(X^*|\boldsymbol{\theta}^*)$)
- **Estimated**: Hat ($\hat{}$) (e.g., $\hat{\boldsymbol{\theta}}$, $\hat{f}(X|\hat{\boldsymbol{\theta}})$)

1 Exercise 1: Laplace Distribution MLE Simulation

The Laplace distribution (also known as the double exponential distribution) is fundamental for robust statistical inference and sparse modeling. It has heavier tails than the normal distribution and is widely used in Bayesian statistics (Laplace prior) and machine learning (L1 regularization). Unlike the normal distribution, it has sharp peak at the mode and exponential decay in both directions.

1.1 Part (a): Mathematical Setup

Task: Write down the complete mathematical setup for a Laplace distribution MLE simulation study using the clean notation system.

Requirements:

1. Define the true data generating process for a Laplace distribution with parameters μ (location) and b (scale)
2. Write the probability density function in mathematical form

3. Specify the observed data notation for T samples
4. Define the candidate model specification using asterisk notation
5. Write the likelihood function and log-likelihood function
6. Derive the MLE optimization problem
7. Specify the estimated model notation after optimization

Hints:

- The Laplace density is:
$$p(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$
where $\mu \in \mathbb{R}$ is the location parameter and $b > 0$ is the scale parameter.
- The MLE estimators for Laplace distribution are:
$$\hat{\mu} = \text{median}(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_T) \quad \text{(sample median)} \tag{1}$$
$$\hat{b} = \frac{1}{T}\sum_{i=1}^{T} |\tilde{X}_i - \hat{\mu}| \quad \text{(mean absolute deviation from median)} \tag{2}$$
- Key properties:
  - $\mathbb{E}[X] = \mu$ and $\text{Var}[X] = 2b^2$
  - The Laplace distribution is symmetric around μ
  - It has exponential tails: $P(|X - \mu| > t) = \exp(-t/b)$

1.2 Part (b): Jupyter Notebook Implementation

Task: Create a complete Jupyter notebook implementing the Laplace MLE simulation.

Requirements:

1. Import necessary libraries (numpy, scipy, matplotlib, pandas)
2. Implement the true data generation process
3. Implement MLE estimation (you may use scipy.optimize or analytical solutions where available)
4. Create the simulation loop with proper result storage
5. Generate summary statistics and visualizations
6. Include proper documentation and comments using the clean notation

1.3 Part (c): Scale Parameter Analysis

Task: Investigate how the scale parameter b affects estimation uncertainty and robustness properties.

Requirements:

1. Run simulations for $b = 0.5, 1.0, 2.0$ (small vs. larger scale values)
2. Fix location parameter: $\mu_{\text{true}} = 0$

3. Use sample size $n = 100$ and $K = 1000$ simulations

4. For each b, analyze:
   - Bias in parameter estimates
   - Standard errors of estimates
   - Distribution shape of estimates
   - Robustness to outliers (add some extreme values)

5. Create comparative visualizations showing how b affects:
   - Parameter estimate distributions
   - Uncertainty quantification quality
   - Median vs. mean performance

Questions to address:

- How does the scale parameter affect estimation uncertainty?
- Why does the Laplace distribution use the median instead of the mean for location estimation?
- How robust is the Laplace MLE to outliers compared to normal distribution MLE?

1.4 Part (e): Sample Size Analysis

Task: Examine how sample size n affects parameter estimation for different scale parameters.
Requirements:

1. Fix b values at 0.5, 1.0, and 2.0

2. Vary sample sizes: $n = 20, 50, 100, 200, 500$

3. Run $K = 1000$ simulations for each combination

4. Analyze how the combination of n and b affects:
   - Bias reduction
   - Standard error scaling
   - Asymptotic behavior of median estimator

5. Create plots showing the relationship between n, b, and estimation quality

Questions to address:

- How does the standard error scaling apply to the median estimator?
- Do larger scale parameters require different sample sizes for reliable estimation?
- How does the finite-sample performance of the median compare to the mean?

2 Exercise 2: Binomial Distribution MLE Simulation

The binomial distribution is fundamental for discrete data analysis. This exercise explores MLE estimation for binomial parameters and their connection to normal approximations.

2.1 Part (a): Mathematical Setup

Task: Write the complete mathematical framework for binomial MLE estimation using clean notation.

Requirements:

1. Define the true binomial data generating process with parameters n (trials) and p (success probability)
2. Write the binomial probability mass function
3. Specify the observed data notation for T independent observations
4. Define the candidate model using asterisk notation
5. Write the likelihood function for binomial data
6. Derive the MLE estimators analytically
7. Specify the estimated model after optimization

Hints:

- For $X \sim \text{Binomial}(n, p)$:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- The MLE estimators for Binomial distribution are:

$$\hat{p} = \frac{1}{T}\sum_{i=1}^{T} \frac{\tilde{X}_i}{n} = \frac{\bar{X}}{n} \quad \text{(sample proportion)} \tag{3}$$

$$\hat{n} = \frac{\bar{X}^2}{\bar{X} - S^2} \quad \text{(method of moments, when } n \text{ unknown)} \tag{4}$$

$$\text{where } \bar{X} = \frac{1}{T}\sum_{i=1}^{T} \tilde{X}_i \text{ and } S^2 = \frac{1}{T-1}\sum_{i=1}^{T}(\tilde{X}_i - \bar{X})^2 \tag{5}$$

- When both n and p are unknown, the method of moments estimators are:

$$\hat{p} = 1 - \frac{S^2}{\bar{X}} \tag{6}$$

$$\hat{n} = \frac{\bar{X}}{1 - S^2/\bar{X}} = \frac{\bar{X}^2}{\bar{X} - S^2} \tag{7}$$

2.2 Part (b): Jupyter Notebook Implementation

Task: Create a comprehensive Jupyter notebook for binomial MLE simulation.

Requirements:

1. Implement true data generation from binomial distribution
2. Implement MLE estimation for binomial parameters
3. Create simulation loops with result collection
4. Generate summary statistics and visualizations
5. Analyze bias and variance properties of estimators
6. Compare different combinations of n and p values

2.3 Part (c): Normal Approximation Analysis

Task: Investigate the normal approximation to the binomial distribution through MLE estimation. Here we assume that true data generating distribution is a binomial with large n but the candidate model for the estimation is a normal distribution.

Requirements:

1. Write down in clean notation the mathematical framework for the normal approximation to the binomial distribution.
2. For the same binomial data, fit both:
   - True binomial MLE: $\hat{n}$, $\hat{p}$
   - Normal approximation MLE: $\hat{\mu}$, $\hat{\sigma}$
3. Use the relationships:

$$\mathbb{E}[X] = np = \mu \tag{8}$$

$$\mathrm{Var}[X] = np(1-p) = \sigma^2 \tag{9}$$

4. Solve for n and p from normal estimates:

$$\hat{n}_{\text{normal}} = \frac{\hat{\mu}^2}{\hat{\mu} - \hat{\sigma}^2} \tag{10}$$

$$\hat{p}_{\text{normal}} = \frac{\hat{\mu} - \hat{\sigma}^2}{\hat{\mu}} \tag{11}$$

5. Compare the two approaches across different parameter combinations

Analysis Requirements:

1. Test parameter combinations:
   - Small n, moderate p: $(n = 10, p = 0.5)$
   - Large n, moderate p: $(n = 100, p = 0.5)$
   - Large n, extreme p: $(n = 100, p = 0.05)$
   - Large n, extreme p: $(n = 100, p = 0.95)$
2. For each combination, analyze:
   - Accuracy of normal approximation to binomial MLE
   - Uncertainty quantification comparison
   - When the normal approximation breaks down
3. Create visualizations comparing:
   - Parameter estimate distributions
   - Standard error estimates
   - Bias and variance properties

Questions to address:

- Under what conditions is the normal approximation reliable?
- How do extreme values of p (near 0 or 1) affect the approximation?
- What is the practical value of the normal approximation for uncertainty quantification?

Deliverables

Submit the following:

1. **Mathematical derivations**: Complete solutions to parts 1(a) and 2(a) with clean notation
2. **Jupyter notebooks**: Well-documented implementations for both exercises
3. **Results report**: Summary of findings including:
   - Key insights from parameter variation studies
   - Comparative analysis across distributions
   - Visualizations with proper interpretations
   - Connection to broader ML uncertainty quantification
4. **Code documentation**: Clear comments using the notation system from lectures