DS-GA 3001.009 Applied Statistics: Homework #6

Due on Thursday, November 16, 2023

Please hand in your homework via Gradescope (entry code: RKXJN2) before 11:59 PM.

1. Revisit the example of bivariate Gaussian location model we covered in class:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \cdots, \begin{bmatrix} x_n \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \theta_0 \\ \eta_0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right),$$

   where $\rho \in [-1, 1]$ is known.

   (a) Recall that the estimating equation based on the score for θ_0 is

$$\frac{1}{n}\sum_{i=1}^{n} \left[x_i - \widehat{\theta} - \rho(y_i - \widehat{\eta}) \right] = 0.$$

   If $\widehat{\eta} = \eta_0$ is the true nuisance, from the above equation, determine the probability distribution of $\widehat{\theta} - \theta_0$ which only depends on (n, ρ).

   (b) Repeat (a) if $\widehat{\eta} = \eta_0 + \varepsilon$ with a fixed constant ε. Your answer should depend on (n, ρ, ε).

   (c) Now consider the efficient score equation

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\theta}) = 0.$$

   Write out the probability distribution of $\widehat{\theta} - \theta_0$. How does $\mathbb{E}[(\widehat{\theta} - \theta_0)^2]$ compare with (a) and (b)?

2. In this problem, we consider a simple error-in-variable model

$$y = \theta_0 z_0 + \varepsilon_1, \quad \varepsilon_1 \sim \mathcal{N}(0, 1),$$
$$x = z_0 + \varepsilon_2, \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma^2).$$

   Here the observables are (x, y), the target parameter is θ_0, the nuisance parameter is z_0, and the errors $(\varepsilon_1, \varepsilon_2)$ are independent. The parameter σ is known.

   (a) Write out the log-likelihood of (x, y) given (θ_0, z_0), up to an additive constant.

   (b) Compute the score functions $s^{\theta}_{(\theta_0, z_0)}(x, y)$ and $s^{z}_{(\theta_0, z_0)}(x, y)$.

   (c) Compute the efficient score function $s^{\text{eff}}_{(\theta_0, z_0)}(x, y)$ for θ_0.

   (d) Now suppose that we have n i.i.d. observations $(x_1, y_1), \cdots, (x_n, y_n)$, as well as a nuisance estimate $\widehat{z}$. Find the estimator $\widehat{\theta}$ based on the efficient score function.

3. Coding: we will implement Stein's semiparametric estimator for the symmetric location model $y_1, \cdots, y_n \sim f(y - \theta_0)$, where in our experiment $f(y) = e^{-|y|}/2$ is the Laplace density. We will experiment on three estimators of θ_0:

- the sample mean of $(y_1, \cdots, y_n)$;
- the MLE with the knowledge of f - you should derive the form of the MLE here and find it to be a very simple statistic of $(y_1, \cdots, y_n)$;
- Stein's semiparametric estimator without the knowledge of f.

Based on inline instructions, fill in the missing codes in `https://tinyurl.com/5zjf4bzd`. Be sure to submit a pdf with your codes, outputs, and colab link.