Stats 210A, Fall 2023
Homework 11

Due date: Wednesday, Nov. 15

1. Some Maximum Likelihood Estimators

Find the MLE for each model below, show that it is consistent, and find its asymptotic distribution. You may assume our Taylor expansions from class are valid without checking conditions.

(a) Binomial: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Binom}(m, \theta)$. Find the MLE for θ and for the natural parameter $\eta = \log \frac{\theta}{1-\theta}$.

(b) Gaussian: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$. Find (i) the MLE for θ if σ^2 is known, (ii) the MLE for σ^2 if θ is known, and (iii) the MLE for (θ, σ^2) if neither is known.

(c) Laplace: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \frac{1}{2}e^{-|x-\theta|}$. Assume n is odd.
For this problem, the log-likelihood is non-differentiable at one point, but we can still use our formula for the asymptotic distribution of the MLE from class, with the Fisher information defined by $J_1(\theta) = \text{Var}_\theta[\dot{\ell}_1(\theta; X_i)]$. You may use this fact without proof.

(d) **Optional** (not graded, no extra points) For the Laplace, plot a few realizations of the log-likelihood for $n = 5000$ with $\theta_0 = 0$, and plot over it the quadratic approximation given by

$$\ell_n(\theta) - \ell_n(\theta_0) \approx \dot{\ell}_n(\theta_0)(\theta - \theta_0) - \frac{1}{2}nJ_1(\theta_0)(\theta - \theta_0)^2.$$

Is the quadratic approximation pretty good in the neighborhood $\theta_0 \pm 3\sigma$, where σ^2 is the approximate variance of $\hat{\theta}_n$? Intuitively, what do you think might account for this when the second derivative doesn't exist?

2. Estimating the inverse of a mean

Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, and that we are interested in estimating the quantity $1/\theta$. In order to do so, we use the estimator $\delta(X) = 1/\overline{X}_n$ where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. Assume $\theta \neq 0$.

(a) Show that δ is asymptotically normal, and find its asymptotic distribution.

(b) Show that the expectation $\mathbb{E}|1/\overline{X}_n| = \infty$ for every n. Why does this not contradict the result of part (a)?

(c) Simulate to find the distribution of $1/\overline{X}_n$ for $n = 10, 100, 10^4$ and $\theta = 0.1, 1, 10$. For each setting of the parameters, plot a histogram of the estimator and overlay its Gaussian approximation. When the Gaussian approximation is not good, what is going wrong? Is the sample size a reliable indicator of whether we should trust an asymptotic approximation?
Hint: If you are using R, the functions `hist` (with argument `freq = FALSE` to get a density histogram), `curve`, and `dnorm` will come in handy. Also, I recommend manually setting the `breaks` and `xlim` arguments in `hist` to stop enormous values from making your histogram uninformative: $\mu \pm 4\sigma$ is a reasonable range of values to plot, where μ and σ^2 are the mean and variance of the Gaussian approximation.

3. Limiting distribution of U-statistics

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$ in some sample space $\mathcal{X}$. $U_n = U_n(X_1, \dots, X_n)$ is called a rank-2 U-statistic if

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} h(X_i, X_j)$$

where h is a symmetric function, i.e. $h(x_1, x_2) = h(x_2, x_1)$ for any $x_1, x_2 \in \mathcal{X}$.

In this problem, we denote $\theta = \mathbb{E}h(X_1, X_2)$ and assume that $\mathbb{E}h(X_1, X_2)^2 < \infty$. Note that U_n is the nonparametric UMVU estimator of θ.

Perhaps surprisingly, we can derive the asymptotic distribution of U_n in a relatively small number of steps using a technique called *Hájek projection* where we approximate it by an additive function of the independent X_i variables. We walk through the proof below.

(a) Define $g(x) = \mathbb{E}h(x, X_2) - \theta = \int h(x, u)\, \mathrm{d}P(u) - \theta$. Show that, for all i,

$$\mathbb{E}g(X_i) = 0, \quad \text{and} \quad \mathrm{Var}(g(X_i)) < \infty.$$

(**Note:** g is a specific function from $\mathcal{X}$ to $\mathbb{R}$. It is not a rule for naively substituting symbols into expressions. In particular, note that $g(X_i)$, a random variable, is not the same as the deterministic expression $\mathbb{E}h(X_i, X_2) - \theta$.)

(b) Define $\widehat{U}_n = \theta + \frac{2}{n} \sum_{i=1}^{n} g(X_i)$. Show that $\mathbb{E}[(U_n - \widehat{U}_n) f(X_i)] = 0$ for any i and any measurable function $f(X_i)$ with $\mathbb{E}[f(X_i)^2] < \infty$.
(**Hint:** Condition on X_i)

(c) Show that $\sqrt{n}(U_n - \widehat{U}_n) \xrightarrow{p} 0$ as $n \to \infty$. (Hint: show that U_n and $\widehat{U}_n$ have the same asymptotic variance, and then apply part (b)).

(d) Conclude that $\sqrt{n}(U_n - \theta) \Rightarrow N(0, 4\zeta_1)$, where $\zeta_1 = \mathrm{Var}(g(X_1))$.

(e) Assume that $\mathcal{X} = \mathbb{R}$ with $\mathbb{E}X_i^4 < \infty$. Express the sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ as a rank-2 U-statistic and use the above results to derive its asymptotic distribution.

(**Note:** a similar result holds in general for rank-r U-statistics if we set $\widehat{U}_n = \theta + \frac{r}{n} \sum_i g(X_i)$ where $g(x) = \mathbb{E}[h(x, X_2, \dots, X_r)] - \theta$.)

Moral: If P^n is the distribution of $(X_1, \dots, X_n)$ then it is easy to check that the set of all square-integrable random variables of the form $f(X_1, \dots, X_n)$ (where $f : \mathcal{X}^n \to \mathbb{R}$ is measurable) forms a vector space over $\mathbb{R}$, which we call $L^2(P^n)$, where we can define an inner product as

$$\langle f(X), g(X) \rangle_{L^2} = \mathbb{E}[f(X) g(X)] \leq \sqrt{\mathbb{E}[f(X)^2] \mathbb{E}[g(X)^2]} < \infty.$$

Moreover, the subset of those random variables that can be written as $\sum_i f_i(X_i)$, where each f_i is measurable, forms a subspace. Part (b) establishes that the simpler random variable $\widehat{U}_n$ is the *projection* of U_n onto this subspace, and part (c) establishes that U_n is asymptotically very close to its projection.

4. Probabilistic big-O notation

Let $X_1, X_2, \dots$ denote a sequence of random vectors (with $\|X_n\| < \infty$ almost surely for each n). We say the sequence is *bounded in probability* (or sometimes *tight*) if for every $\varepsilon > 0$ there exists a constant $M_\varepsilon > 0$ for which

$$\mathbb{P}(\|X_n\| > M_\varepsilon) < \varepsilon, \quad \forall n.$$

Informally, there is "no mass escaping to infinity" as n grows. Like regular big-O notation, these symbols can help to make rigorous asymptotic proofs look clean and intuitive.

For a fixed sequence a_n, we say $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$ as $n \to \infty$, and $X_n = O_p(a_n)$ if the sequence $(X_n/a_n)_{n\geq 1}$ is bounded in probability.

Prove the following facts for $X_n, Y_n \in \mathbb{R}^d$:

(a) If $X_n \Rightarrow X$ for any random vector X, then $X_n = O_p(1)$.

(b) If $X_n = o_p(a_n)$ then $X_n = O_p(a_n)$.

(c) If $a_n/b_n \to 0$ and $X_n = O_p(a_n)$, then $X_n = o_p(b_n)$.

(d) If $X_n = O_p(a_n)$ and $Y_n = O_p(b_n)$ then $X_n + Y_n = O_p(\max\{a_n, b_n\})$.

(e) If $X_n = O_p(a_n)$ and $Y_n = o_p(b_n)$, then $X_n' Y_n = o_p(a_n b_n)$. If $X_n = O_p(a_n)$ and $Y_n = O_p(b_n)$, then $X_n' Y_n = O_p(a_n b_n)$.

(f) If $X_n = O_p(1)$ and $g : \mathbb{R}^d \to \mathbb{R}^k$ is continuous then $g(X_n) = O_p(1)$.

(g) For $d = 1$, if $X_n = O_p(a_n)$ with $a_n \to 0$ and $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable with $g(0) = \dot{g}(0) = 0$, then $g(X_n) = o_p(a_n)$. Show further that if g is twice continuously differentiable then $g(X_n) = O_p(a_n^2)$. (**Hint:** Use the mean value theorem and apply a previous part of this problem.)

(h) For $d = 1$, if $\mathrm{Var}(X_n) = a_n^2 < \infty$ and $\mathbb{E}X_n = b_n$ then $X_n = O_p(a_n + b_n)$. (**Hint:** Use Chebyshev's inequality.)

(i) **Optional** (not graded, no extra points): If $\mathrm{Var}(X_n) = a_n^2 < \infty$, is it impossible to have $X_n = o_p(a_n)$? Prove or give a counterexample.