Stats 210A, Fall 2023
Homework 12

Optional

1. MLE consistency for concave log-likelihoods

Assume $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}(x)$ for some identifiable, dominated family with $\Theta = \mathbb{R}^d$. Assume additionally that $\ell_1(\theta; X_i) = \log p_\theta(X_i)$ is almost surely concave and continuously differentiable in θ, and that for all compact sets $K \subseteq \mathbb{R}^d$, we have

$$\mathbb{E}_{\theta_0}\left[\sup_{\theta \in K} \|\nabla \ell_1(\theta; X_1)\|_2\right] < \infty.$$

Prove that the MLE is consistent: if $\hat{\theta}_n \in \text{argmax}\ \ell_n(\theta)$ then $\hat{\theta}_n \xrightarrow{p} \theta_0$ (You may assume a maximizer always exists; note we could always define $\hat{\theta}_n$ arbitrarily when there is none).

(**Hint:** The technique here is not just a small modification of what we used in our theorem from class for consistency with non-compact Θ; it's a different argument entirely. But similarly to what we did in class, you should start by showing uniform convergence of $\overline{W}_n(\theta)$ on compact K, and then deal with the rest of $\mathbb{R}^d$.)

Moral: There is more than one way to get consistency of the MLE.

2. Logistic regression with random X

Consider a univariate logistic regression model where we observe n i.i.d. pairs $(X_i, Y_i) \in \mathbb{R} \times \{0, 1\}$. The covariate is random with a known distribution, $X_i \overset{\text{i.i.d.}}{\sim} U[-1, 1]$, and

$$\mathbb{P}_{\alpha,\beta}(Y_i = 1 \mid X_i = x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

(a) Show that the maximum likelihood estimator for (α, β) solves

$$\sum_i Y_i = \sum_i \pi_i(\hat{\alpha}_n, \hat{\beta}_n)$$
$$\sum_i Y_i X_i = \sum_i \pi_i(\hat{\alpha}_n, \hat{\beta}_n) X_i,$$

where $\pi_i(\alpha, \beta) = e^{\alpha+\beta X_i}/(1 + e^{\alpha+\beta X_i})$.

(b) Use the result of the previous problem to show that the MLE is consistent, asymptotically Gaussian, and asymptotically efficient (you may ignore the fact that the MLE may not always exist in finite samples).

(c) For $\alpha = 0$, $\beta = 4$, calculate the Fisher information for a single pair (X_i, Y_i); give it as an integral and also calculate it numerically (you do not need to analytically evaluate the integral). Note your answer should not depend on X_i, which is a random variable in this problem. Give the asymptotic distribution of the MLE, with a numerical answer for the asymptotic variance.

(d) For $\alpha = 0$, $\beta = 4$, and for each of a few different n values:

(i) Generate a large number (e.g. 1000) of data sets of size n, and for each one compute the MLE $(\hat{\alpha}, \hat{\beta})$ (you can use statistical software to compute the MLE, e.g. the `glm` function in R).

(ii) Plot histograms of $\hat{\alpha}$ and $\hat{\beta}$ (if you use R, I recommend setting `freq=FALSE` to get a density histogram instead of a frequency histogram).

(iii) Overlay the Gaussian curve based on the approximate distribution from part (c) (you can use the `dnorm` function in R). About how big does n need to be for the normal approximation to be pretty good?

(e) Repeat parts (c) and (d) for $\alpha = -6$ and $\beta = 4$. How is it the same or different, and what do you think accounts for why?

3. Score test with nuisance parameters

Consider a testing problem with $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta,\zeta}(x)$ with parameter of interest $\theta \in \mathbb{R}$ and nuisance parameter $\zeta \in \mathbb{R}$. That is, we are testing $H_0 : \theta = \theta_0$ vs. $H_1 : \ \theta \neq \theta_0$, and ζ is unknown; let ζ_0 denote its true value. Then there is a version of the score test where we plug in an estimator for ζ, but we must use a corrected version of the variance.

Let $\hat{\zeta}_0$ denote the maximum likelihood estimator of ζ under the null:

$$\hat{\zeta}_0(\theta_0) = \arg\max_{\zeta \in \mathbb{R}} \ \ \ell(\theta_0, \zeta; X).$$

Assume $\hat{\zeta}_0$ is consistent under the null hypothesis.

Let $J(\theta, \zeta)$ denote the full-sample Fisher Information (omitting the usual n subscript), and assume it is continuous and positive-definite everywhere.

(a) Use Taylor expansions informally to show that, for large n,

$$\frac{\partial}{\partial \theta}\ell(\theta_0, \hat{\zeta}_0) \approx \frac{\partial}{\partial \theta}\ell(\theta_0, \zeta_0) - \frac{\frac{\partial^2}{\partial\theta\partial\zeta}\ell(\theta_0, \zeta_0)}{\frac{\partial^2}{\partial\zeta^2}\ell(\theta_0, \zeta_0)} \ \frac{\partial}{\partial \zeta}\ell(\theta_0, \zeta_0).$$

(Note: the LHS should be read as $[\frac{\partial}{\partial\theta}\ell(\theta, \zeta)]\big|_{\theta_0, \hat{\zeta}_0}$, and **not** $\frac{d}{d\theta_0}[\ell(\theta_0, \hat{\zeta}_0(\theta_0))]$).

(b) Using part (a), conclude that

$$\left(J_{11} - \frac{J_{12}^2}{J_{22}}\right)^{-1/2} \frac{\partial}{\partial \theta}\ell(\theta_0, \hat{\zeta}_0) \Rightarrow N(0,1) \quad \text{ as } n \to \infty$$

where $J = J(\theta_0, \hat{\zeta}_0)$. Compare this to the score test statistic we would use if ζ_0 were known rather than estimated. (Note: you may assume without proof that the approximation error in part (a) is negligible; i.e. you may take the "$\approx$" as an exact equality).

Moral: The score test can be carried out with nuisance parameters, but the fact that we estimate the nuisance parameter affects the distribution of the test statistic in a way that we need to take into account.

4. Poisson score test

Suppose that for $i = 1, \ldots, x_n$ we observe a real covariate $x_i \in \mathbb{R}$ (fixed and known) and a Poisson response $Y_i \sim$ Pois(λ_i). We assume that $\lambda_i = \alpha + \beta x_i$, with the restriction that $\lambda_i \geq 0$ for all i, but with $\alpha, \beta \in \mathbb{R}$ otherwise unrestricted. Assume that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n |x_i - \bar{x}_n|^3}{\left(\sum_{i=1}^n (x_i - \bar{x}_n)^2\right)^{3/2}} = 0,$$

where $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$. We observe the first n pairs (x_i, y_i) and our goal is to test the hypothesis $H_0 : \ \beta = 0$ vs. $H_1 : \ \beta > 0$. Assume that there are at least 3 distinct values represented among $x_1, \ldots, x_n$.

(a) Show that this model is a curved exponential family.

(b) Derive the score test statistic for H_0 vs H_1. Give the test statistic and asymptotic rejection cutoff. It is not necessary to normalize it for this part.

(c) Show that your test statistic is indeed asymptotically normally distributed, and find an asymptotically valid rejection cutoff.

Hint: It may help to use the *Lyapunov CLT*, which applies to sums of independent random variables that are not necessarily identically distributed: Suppose $Z_1, Z_2, \ldots$ is a sequence of random variables with $Z_i \sim (\mu_i, \sigma_i^2)$, for $\sigma_i^2 < \infty$. Define $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If for some $\delta > 0$, we have

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}\left[|Z_i - \mu_i|^{2+\delta}\right] = 0,$$

then $s_n^{-1} \sum_{i=1}^n (Z_i - \mu_i) \Rightarrow N(0, 1)$.

(d) Suppose n is small, so we don't want to rely on the asymptotic normality. Explain how we could find a finite-sample exact conditional cutoff for the score test from part (b) (it is not necessary to prove any optimality property).

5. Super-Efficient Estimator

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ and consider estimating θ via:

$$\delta_n(X) = \overline{X}_n 1\{|\overline{X}_n| > a_n\},$$

where $a_n \to 0$ but $a_n\sqrt{n} \to \infty$ as $n \to \infty$ (for example, $a_n = n^{-1/4}$).

(a) Show that δ_n has the same asymptotic distribution as $\overline{X}_n$ when $\theta \neq 0$, but that $\sqrt{n}(\delta_n - 0) \xrightarrow{p} 0$ if $\theta = 0$.

(b) Show that, pointwise in θ, as $n \to \infty$,

$$n\,\text{MSE}(\delta_n; \theta) \to 1\{\theta \neq 0\},$$

but that the convergence is not uniform in θ; in fact,

$$\sup_{\theta\in\mathbb{R}} \; n\,\text{MSE}(\delta_n; \theta) \to \infty.$$

(**Note**: this is an example of a situation where it is incorrect to exchange a limit with a supremum.)

(c) **Optional**: Can you find a scaling of δ_n that converges to a non-degenerate distribution when $\theta = 0$? What is the limiting distribution?

Moral: The sense in which asymptotically efficient estimators are "optimal" is not easy to define, and it isn't obvious how we should compare the asymptotic behavior of different estimators. In this example it would appear initially that the super-efficient estimator renders the sample mean inadmissible. But this is only true if we look at the pointwise limit for fixed θ; at any n there are some values of θ for which the estimator is performing very badly, and this gets worse and worse as n gets larger.