Lecture 16 — MLE under model misspecification

The eminent statistician George Box once said,

"All models are wrong, but some are useful."

When we fit a parametric model to a set of data $X_1, \ldots, X_n$, we are usually not certain that the model is correct (for example, that the data truly have a normal or Gamma distribution). Rather, we think of the model as an approximation to what might be the true distribution of data. It is natural to ask, then, whether the MLE estimate $\hat{\theta}$ in a parametric model is at all meaningful, if the model itself is incorrect. The goal of this lecture is to explore this question and to discuss how the properties of $\hat{\theta}$ change under model misspecification.

16.1 MLE and the KL-divergence

Consider a parametric model $\{f(x|\theta) : \theta \in \Omega\}$. We'll assume throughout this lecture that $f(x|\theta)$ is the PDF of a continuous distribution, and $\theta \in \mathbb{R}$ is a single parameter.

Thus far we have been measuring the error of an estimator $\hat{\theta}$ by its distance to the true parameter θ, via the bias, variance, and MSE. If $X_1, \ldots, X_n \overset{IID}{\sim} g$ for a PDF g that is not in the model, then there is no true parameter value θ associated to g. We will instead think about a measure of "distance" between two general PDFs:

Definition 16.1. For two PDFs f and g, the **Kullback-Leibler (KL) divergence** from f to g is

$$D_{\mathrm{KL}}(g\|f) = \int g(x) \log \frac{g(x)}{f(x)}\, dx.$$

Equivalently, if $X \sim g$, then

$$D_{\mathrm{KL}}(g\|f) = \mathbb{E}\left[\log \frac{g(X)}{f(X)}\right].$$

D_{KL} has many information-theoretic interpretations and applications. For our purposes, we'll just note the following properties: If $f = g$, then $\log(g(x)/f(x)) \equiv 0$, so $D_{\mathrm{KL}}(g\|f) = 0$. By Jensen's inequality, since $x \mapsto -\log x$ is convex,

$$D_{\mathrm{KL}}(g\|f) = \mathbb{E}\left[-\log \frac{f(X)}{g(X)}\right] \geq -\log \mathbb{E}\left[\frac{f(X)}{g(X)}\right] = -\log \int \frac{f(x)}{g(x)} g(x) dx = 0.$$

Furthermore, since $x \mapsto -\log x$ is strictly convex, the inequality above can only be an equality if $f(X)/g(X)$ is a constant random-variable, so $f = g$. Thus, like an ordinary distance measure, $D_{\mathrm{KL}}(g\|f) \geq 0$ always, and $D_{\mathrm{KL}}(g\|f) = 0$ if and only if $f = g$.

Example 16.2. To get an intuition for what the KL-divergence is measuring, let f and g be the PDFs of the distributions $\mathcal{N}(\mu_0, \sigma^2)$ and $\mathcal{N}(\mu_1, \sigma^2)$. Then

$$\begin{aligned}\log \frac{g(x)}{f(x)} &= \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} \Big/ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} \right) \\ &= -\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_0)^2}{2\sigma^2} \\ &= \frac{2(\mu_1-\mu_0)x - (\mu_1^2 - \mu_0^2)}{2\sigma^2}.\end{aligned}$$

So letting $X \sim g$,

$$\begin{aligned}D_{\mathrm{KL}}(g\|f) = \mathbb{E}\left[\log \frac{g(X)}{f(X)}\right] &= \frac{1}{2\sigma^2}\left(2(\mu_1-\mu_0)\mathbb{E}[X] - (\mu_1^2-\mu_0^2)\right) \\ &= \frac{1}{2\sigma^2}\left(2(\mu_1-\mu_0)\mu_1 - (\mu_1^2-\mu_0^2)\right) = \frac{(\mu_1-\mu_0)^2}{2\sigma^2}.\end{aligned}$$

Thus $D_{\mathrm{KL}}(g\|f)$ is proportional to the square of the mean difference normalized by the standard deviation σ. In this example we happen to have $D_{\mathrm{KL}}(f\|g) = D_{\mathrm{KL}}(g\|f)$, but in general this is not true—for two arbitrary PDFs f and g, we may have $D_{\mathrm{KL}}(f\|g) \neq D_{\mathrm{KL}}(g\|f)$.

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} g$, and consider a parametric model $\{f(x|\theta) : \theta \in \Omega\}$ which may or may not contain the true PDF g. The MLE $\hat{\theta}$ is the value of θ that maximizes

$$\frac{1}{n} l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta),$$

and this quantity by the Law of Large Numbers converges in probability to

$$\mathbb{E}_g[\log f(X|\theta)]$$

where $\mathbb{E}_g$ denotes expectation with respect to $X \sim g$. In Lecture 14, we showed that when $g(x) = f(x|\theta_0)$ (meaning g belongs to the parametric model, and the true parameter is θ_0), then $\mathbb{E}_g[\log f(X|\theta)]$ is maximized at $\theta = \theta_0$—this explained consistency of the MLE. More generally, when g does not necessarily belong to the parametric model, we may write

$$\mathbb{E}_g[\log f(X|\theta)] = \mathbb{E}_g[\log g(X)] - \mathbb{E}_g\left[\log \frac{g(X)}{f(X|\theta)}\right] = \mathbb{E}_g[\log g(X)] - D_{\mathrm{KL}}(g\|f(x|\theta)).$$

The term $\mathbb{E}_g[\log g(X)]$ does not depend on θ, so the value of θ maximizing $\mathbb{E}_g[\log f(X|\theta)]$ is the value of θ that minimizes $D_{\mathrm{KL}}(g\|f(x|\theta))$. This (heuristically) shows the following result:[1]

[1]For a rigorous statement of necessary regularity conditions, see for example Halbert White (1982) "Maximum likelihood estimation of misspecified models".

Theorem 16.3. *Let $X_1, \ldots, X_n \overset{IID}{\sim} g$ and suppose $D_{\mathrm{KL}}(g \| f(x|\theta))$ has a unique minimum at $\theta = \theta^*$. Then, under suitable regularity conditions on $\{f(x|\theta) : \theta \in \Omega\}$ and on g, the MLE $\hat{\theta}$ converges to θ^* in probability as $n \to \infty$.*

The density $f(x|\theta^*)$ may be interpreted as the "KL-projection" of g onto the parametric model $\{f(x|\theta) : \theta \in \Omega\}$. In other words, the MLE is estimating the distribution in our model that is closest, with respect to KL-divergence, to g.

16.2 The sandwich estimator of variance

When $X_1, \ldots, X_n \overset{IID}{\sim} g$, how close is the MLE $\hat{\theta}$ to this KL-projection θ^*? Analogous to our proof in Lecture 14, we may answer this question by performing a Taylor expansion of the identity $0 = l'(\hat{\theta})$ around the point $\hat{\theta} = \theta^*$. This yields

$$0 \approx l'(\theta^*) + (\hat{\theta} - \theta^*) l''(\theta^*),$$

so

$$\sqrt{n}(\hat{\theta} - \theta^*) \approx -\frac{\frac{1}{\sqrt{n}} l'(\theta^*)}{\frac{1}{n} l''(\theta^*)}. \tag{16.1}$$

Recall the score function

$$z(x, \theta) = \frac{\partial}{\partial \theta} \log f(x|\theta).$$

The Law of Large Numbers applied to the denominator of (16.1) gives

$$\frac{1}{n} l''(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} z'(X_i, \theta^*) \to \mathbb{E}_g[z'(X, \theta^*)]$$

in probability, while the Central Limit Theorem applied to the numerator of (16.1) gives

$$\frac{1}{\sqrt{n}} l'(\theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z(X_i, \theta^*) \to \mathcal{N}(0, \mathrm{Var}_g[z(X, \theta^*)])$$

in distribution. (The quantity $z(X, \theta^*)$ has mean 0 when $X \sim g$ because θ^* maximizes $\mathbb{E}_g[\log f(X|\theta)]$, so differentiating with respect to θ yields

$$0 = \mathbb{E}_g \left[\frac{\partial}{\partial \theta} [\log f(X|\theta)]_{\theta = \theta^*} \right] = \mathbb{E}_g[z(X, \theta^*)].)$$

Hence by Slutsky's lemma,

$$\sqrt{n}(\hat{\theta} - \theta^*) \to \mathcal{N}\left(0, \frac{\mathrm{Var}_g[z(X, \theta^*)]}{\mathbb{E}_g[z'(X, \theta^*)]^2}\right).$$

These are the same formulas as in Lecture 14 (with θ^* in place of θ_0), except expectations and variances are taken with respect to $X \sim g$ rather than $X \sim f(x|\theta^*)$. If $g(x) = f(x|\theta^*)$,

meaning the model is correct, then $\mathrm{Var}_g[z(X,\theta^*)] = -\mathbb{E}_g[z'(X,\theta^*)] = I(\theta^*)$, and we recover our theorem from Lecture 14. However, when $g(x) \neq f(x|\theta^*)$, in general

$$\mathrm{Var}_g[z(X,\theta^*)] \neq -\mathbb{E}_g[z'(X,\theta^*)],$$

so we cannot simplify the variance of the above normal limit any further. We may instead estimate the individual quantities $\mathrm{Var}_g[z(X,\theta^*)]$ and $\mathbb{E}_g[z'(X,\theta^*)]$ using the sample variance of $z(X_i,\hat{\theta})$ and the sample mean of $z'(X_i,\hat{\theta})$—this yields the **sandwich estimator** for the variance of the MLE.

Example 16.4. Suppose we fit the model Exponential(λ) to data $X_1,\ldots,X_n$ by computing the MLE. The log-likelihood is

$$l(\lambda) = \sum_{i=1}^n \log \lambda e^{-\lambda X_i} = n\log\lambda - \lambda\sum_{i=1}^n X_i,$$

so the MLE solves the equation $0 = l'(\lambda) = n/\lambda - \sum_{i=1}^n X_i$. This yields the MLE $\hat{\lambda} = 1/\bar{X}$ (which is the same as the method-of-moments estimator from Lecture 12).

We may compute the sandwich estimate of the variance of $\hat{\lambda}$ as follows: In the exponential model,

$$z(x,\lambda) = \frac{\partial}{\partial\lambda}\log f(x|\lambda) = \frac{1}{\lambda} - x, \qquad z'(x,\lambda) = \frac{\partial^2}{\partial\lambda^2}\log f(x|\lambda) = -\frac{1}{\lambda^2}.$$

Let $\bar{Z} = \frac{1}{n}\sum_{i=1}^n z(X_i,\hat{\lambda}) = \frac{1}{n}\sum_{i=1}^n(\frac{1}{\hat{\lambda}} - X_i) = \frac{1}{\hat{\lambda}} - \bar{X}$ be the sample mean of $z(X_1,\hat{\lambda})$, ..., $z(X_n,\hat{\lambda})$. We estimate $\mathrm{Var}_g[z(X,\lambda)]$ by the sample variance of $z(X_1,\hat{\lambda}),\ldots,z(X_n,\hat{\lambda})$:

$$\frac{1}{n-1}\sum_{i=1}^n (z(X_i,\hat{\lambda}) - \bar{Z})^2 = \frac{1}{n-1}\sum_{i=1}^n \left((\tfrac{1}{\hat{\lambda}} - X_i) - (\tfrac{1}{\hat{\lambda}} - \bar{X})\right)^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = S_X^2.$$

We estimate $\mathbb{E}_g[z'(X,\lambda)]$ by the sample mean of $z'(X_1,\hat{\lambda}),\ldots,z'(X_n,\hat{\lambda})$:

$$\frac{1}{n}\sum_{i=1}^n z'(X_i,\hat{\lambda}) = \frac{1}{n}\sum_{i=1}^n -\frac{1}{\hat{\lambda}^2} = -\frac{1}{\hat{\lambda}^2}.$$

So the sandwich estimate of $\mathrm{Var}_g[z(X,\lambda)]/\mathbb{E}_g[z'(X,\lambda)]^2$ is $S_X^2\hat{\lambda}^4 = S_X^2/\bar{X}^4$, and we may estimate the standard error of $\hat{\lambda}$ by $S_X/(\bar{X}^2\sqrt{n})$.

In Homework 6, you will compare this sandwich estimator to the usual estimator based on the Fisher information, when $X_1,\ldots,X_n$ do not truly come from an exponential model.