STATS 200: Introduction to Statistical Inference **Autumn 2016**

Lecture 15 — Fisher information and the Cramer-Rao bound

15.1 Fisher information for one or more parameters

For a parametric model $\{f(x|\theta) : \theta \in \Omega\}$ where $\theta \in \mathbb{R}$ is a single parameter, we showed last lecture that the MLE $\hat{\theta}_n$ based on $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$ is, under certain regularity conditions, asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

in distribution as $n \to \infty$, where

$$I(\theta) := \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

is the **Fisher information**. As an application of this result, let us study the sampling distribution of the MLE in a one-parameter Gamma model:

Example 15.1. Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, 1)$. (For this example, we are assuming that we know $\beta = 1$ and only need to estimate α.) Then

$$\log f(x|\alpha) = \log\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x} = -\log\Gamma(\alpha) + (\alpha-1)\log x - x.$$

The log-likelihood of all observations is then

$$\begin{aligned} l(\alpha) &= \sum_{i=1}^n (-\log\Gamma(\alpha) + (\alpha-1)\log X_i - X_i) \\ &= -n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i. \end{aligned}$$

Introducing the digamma function $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$, the MLE $\hat{\alpha}$ is obtained by (numerically) solving

$$0 = l'(\alpha) = -n\psi(\alpha) + \sum_{i=1}^n \log X_i.$$

What is the sampling distribution of $\hat{\alpha}$? We compute

$$\frac{\partial^2}{\partial\alpha^2}\log f(x|\alpha) = -\psi'(\alpha).$$

As this does not depend on x, the Fisher information is $I(\alpha) = -\mathbb{E}_\alpha[-\psi'(\alpha)] = \psi'(\alpha)$. Then for large n, $\hat{\alpha}$ is distributed approximately as $\mathcal{N}(\alpha, \frac{1}{n\psi'(\alpha)})$.

Asymptotic normality of the MLE extends naturally to the setting of multiple parameters:

Theorem 15.2. *Let $\{f(x|\theta) : \theta \in \Omega\}$ be a parametric model, where $\theta \in \mathbb{R}^k$ has k parameters. Let $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$ for $\theta \in \Omega$, and let $\hat{\theta}_n$ be the MLE based on $X_1, \ldots, X_n$. Define the* **Fisher information matrix** *$I(\theta) \in \mathbb{R}^{k \times k}$ as the matrix whose (i, j) entry is given by the equivalent expressions*

$$I(\theta)_{ij} = \operatorname{Cov}_\theta \left[\frac{\partial}{\partial \theta_i} \log f(X|\theta), \frac{\partial}{\partial \theta_j} \log f(X|\theta)\right] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X|\theta)\right]. \tag{15.1}$$

Then under the same conditions as Theorem 14.1,

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, I(\theta)^{-1}),$$

where $I(\theta)^{-1}$ is the $k \times k$ matrix inverse of $I(\theta)$ (and the distribution on the right is the multivariate normal distribution having this covariance).

(For $k = 1$, this definition of $I(\theta)$ is exactly the same as our previous definition, and $I(\theta)^{-1}$ is just $\frac{1}{I(\theta)}$. The proof of the above result is analogous to the $k = 1$ case from last lecture, employing a multivariate Taylor expansion of the equation $0 = \nabla l(\hat{\theta})$ around $\hat{\theta} = \theta_0$.)

Example 15.3. Consider now the full Gamma model, $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$. Numerical computation of the MLEs $\hat{\alpha}$ and $\hat{\beta}$ in this model was discussed in Lecture 13. To approximate their sampling distributions, note

$$\log f(x|\alpha, \beta) = \log \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log x - \beta x,$$

so

$$\frac{\partial^2}{\partial \alpha^2} \log f(x|\alpha, \beta) = -\psi'(\alpha), \quad \frac{\partial^2}{\partial \alpha \partial \beta} \log f(x|\alpha, \beta) = \frac{1}{\beta}, \quad \frac{\partial^2}{\partial \beta^2} \log f(x|\alpha, \beta) = -\frac{\alpha}{\beta^2}.$$

These partial derivatives again do not depend on x, so the Fisher information matrix is

$$I(\alpha, \beta) = \begin{pmatrix} \psi'(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix},$$

and its inverse is

$$I(\alpha, \beta)^{-1} = \frac{1}{\psi'(\alpha)\frac{\alpha}{\beta^2} - \frac{1}{\beta^2}} \begin{pmatrix} \frac{\alpha}{\beta^2} & \frac{1}{\beta} \\ \frac{1}{\beta} & \psi'(\alpha) \end{pmatrix}.$$

$(\hat{\alpha}, \hat{\beta})$ is approximately distributed as the bivariate normal distribution $\mathcal{N}\left((\alpha, \beta), \frac{1}{n} I(\alpha, \beta)^{-1}\right)$. In particular, the marginal distribution of $\hat{\alpha}$ is approximately

$$\mathcal{N}\left(\alpha, \frac{1}{n(\psi'(\alpha)\frac{\alpha}{\beta^2} - \frac{1}{\beta^2})} \frac{\alpha}{\beta^2}\right).$$

Suppose, in this example, that in fact the true parameter $\beta = 1$. Then the variance of $\hat{\alpha}$ reduces to $\frac{1}{n(\psi'(\alpha) - 1/\alpha)}$, which is *not* the variance $\frac{1}{n\psi'(\alpha)}$ obtained in Example 15.1—the variance here is larger. The difference is that in this example, we do not assume that we know $\beta = 1$ and instead are estimating β by its MLE $\hat{\beta}$. As a result, the MLEs of α in these two examples are not the same, and here our uncertainty about β is also increasing the variability of our estimate of α.

More generally, for any 2 × 2 Fisher information matrix

$$I = \begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

the first definition of equation (15.1) implies that $a, c \geq 0$. The upper-left element of I^{-1} is $\frac{1}{a-b^2/c}$, which is always at least $\frac{1}{a}$. This implies, for any model with a single parameter θ_1 that is contained inside a larger model with parameters (θ_1, θ_2), that the variability of the MLE for θ_1 in the larger model is always at least that of the MLE for θ_1 in the smaller model; they are equal when the off-diagonal entry b is equal to 0. The same observation is true for any number of parameters $k \geq 2$ in the larger model.

This is a simple example of a trade-off between model complexity and accuracy of estimation, which is fundamental to many areas of statistics and machine learning: a complex model with more parameters might better capture the true distribution of data, but these parameters will also be more difficult to estimate than those in a simpler model.

15.2 The Cramer-Rao lower bound

Let's return to the setting of a single parameter $\theta \in \mathbb{R}$. Why is the Fisher information $I(\theta)$ called "information", and why should we choose to estimate θ by the MLE $\hat{\theta}$?

If $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta_0)$ for a true parameter θ_0, and $l(\theta) = \sum_{i=1}^n \log f(X_i|\theta)$ is the log-likelihood function, then

$$I(\theta_0) = -\mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\Big[\log f(X|\theta)\Big]_{\theta=\theta_0}\right] = -\frac{1}{n}\mathbb{E}_{\theta_0}[l''(\theta_0)].$$

$I(\theta_0)$ measures the expected curvature of the log-likelihood function $l(\theta)$ around the true parameter $\theta = \theta_0$. If $l(\theta)$ is sharply curved around θ_0—in other words, $I(\theta_0)$ is large—then a small change in θ can lead to a large decrease in the log-likelihood $l(\theta)$, and hence the data provides a lot of "information" that the true value of θ is close to θ_0. Conversely, if $I(\theta_0)$ is small, then a small change in θ does not affect $l(\theta)$ by much, and the data provides less information about θ. In this (heuristic) sense, $I(\theta_0)$ quantifies the amount of information that each observation X_i contains about the unknown parameter.

The Fisher information $I(\theta)$ is an intrinsic property of the model $\{f(x|\theta) : \theta \in \Omega\}$, not of any specific estimator. (We've shown that it is related to the variance of the MLE, but its definition does not involve the MLE.) There are various information-theoretic results stating that $I(\theta)$ describes a fundamental limit to how accurate *any* estimator of θ based on $X_1, \ldots, X_n$ can be. We'll prove one such result, called the **Cramer-Rao lower bound**:

Theorem 15.4. *Consider a parametric model $\{f(x|\theta) : \theta \in \Omega\}$ (satisfying certain mild regularity assumptions) where $\theta \in \mathbb{R}$ is a single parameter. Let T be any unbiased estimator of θ based on data $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$. Then*

$$\mathrm{Var}_\theta[T] \geq \frac{1}{nI(\theta)}.$$

Proof. Recall the score function

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

and let $Z := Z(X_1, \ldots, X_n, \theta) = \sum_{i=1}^n z(X_i, \theta)$. By the definition of correlation and the fact that the correlation of two random variables is always between -1 and 1,

$$\mathrm{Cov}_\theta[Z, T]^2 \leq \mathrm{Var}_\theta[Z] \times \mathrm{Var}_\theta[T].$$

The random variables $z(X_1, \theta), \ldots, z(X_n, \theta)$ are IID, and by Lemma 14.1, they have mean 0 and variance $I(\theta)$. Then

$$\mathrm{Var}_\theta[Z] = n \, \mathrm{Var}_\theta[z(X_1, \theta)] = nI(\theta).$$

Since T is unbiased,

$$\theta = \mathbb{E}_\theta[T] = \int_{\mathbb{R}^n} T(x_1, \ldots, x_n) f(x_1|\theta) \times \ldots \times f(x_n|\theta) dx_1 \ldots dx_n.$$

Differentiating both sides with respect to θ and applying the product rule of differentiation,

$$\begin{aligned} 1 = \int_{\mathbb{R}^n} T(x_1, \ldots, x_n) \bigg(& \frac{\partial}{\partial \theta} f(x_1|\theta) \times f(x_2|\theta) \times \ldots \times f(x_n|\theta) \\ & + f(x_1|\theta) \times \frac{\partial}{\partial \theta} f(x_2|\theta) \times \ldots \times f(x_n|\theta) + \ldots \\ & + f(x_1|\theta) \times f(x_2|\theta) \times \ldots \times \frac{\partial}{\partial \theta} f(x_n|\theta) \bigg) dx_1 \ldots dx_n \\ = \int_{\mathbb{R}^n} & T(x_1, \ldots, x_n) Z(x_1, \ldots, x_n, \theta) f(x_1|\theta) \times \ldots \times f(x_n|\theta) dx_1 \ldots dx_n \\ = \mathbb{E}_\theta & [TZ]. \end{aligned}$$

Since $\mathbb{E}_\theta[Z] = 0$, this implies $\mathrm{Cov}_\theta[T, Z] = \mathbb{E}_\theta[TZ] = 1$, so $\mathrm{Var}_\theta[T] \geq \frac{1}{nI(\theta)}$ as desired. □

For two unbiased estimators of θ, the ratio of their variances is called their **relative efficiency**. An unbiased estimator is **efficient** if its variance equals the lower bound $\frac{1}{nI(\theta)}$. Since the MLE achieves this lower bound asymptotically, we say it is **asymptotically efficient**.

The Cramer-Rao bound ensures that no unbiased estimator can achieve asymptotically lower variance than the MLE. Stronger results, which we will not prove in this class, in fact show that no estimator, biased or unbiased, can asymptotically achieve lower mean-squared-error than $\frac{1}{nI(\theta)}$, except possibly on a small set of special values $\theta \in \Omega$.[1] In particular, when the method-of-moments estimator differs from the MLE, we expect it to have higher mean-squared-error than the MLE for large n, which explains why the MLE is usually the preferred estimator in simple parametric models.

[1] For example, the constant estimator $\hat{\theta} = c$ for fixed $c \in \Omega$ achieves 0 mean-squared-error if the true parameter happened to be the special value c, but at all other parameter values is worse than the MLE for sufficiently large n.