S&DS 242/542: Theory of Statistics

Lecture 15: Fisher information and the Cramer-Rao bound

Recap: Asymptotic normality of the MLE

Let $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ with a single parameter $\theta \in \mathbb{R}$. Under regularity conditions, the MLE $\hat{\theta}$ satisfies

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

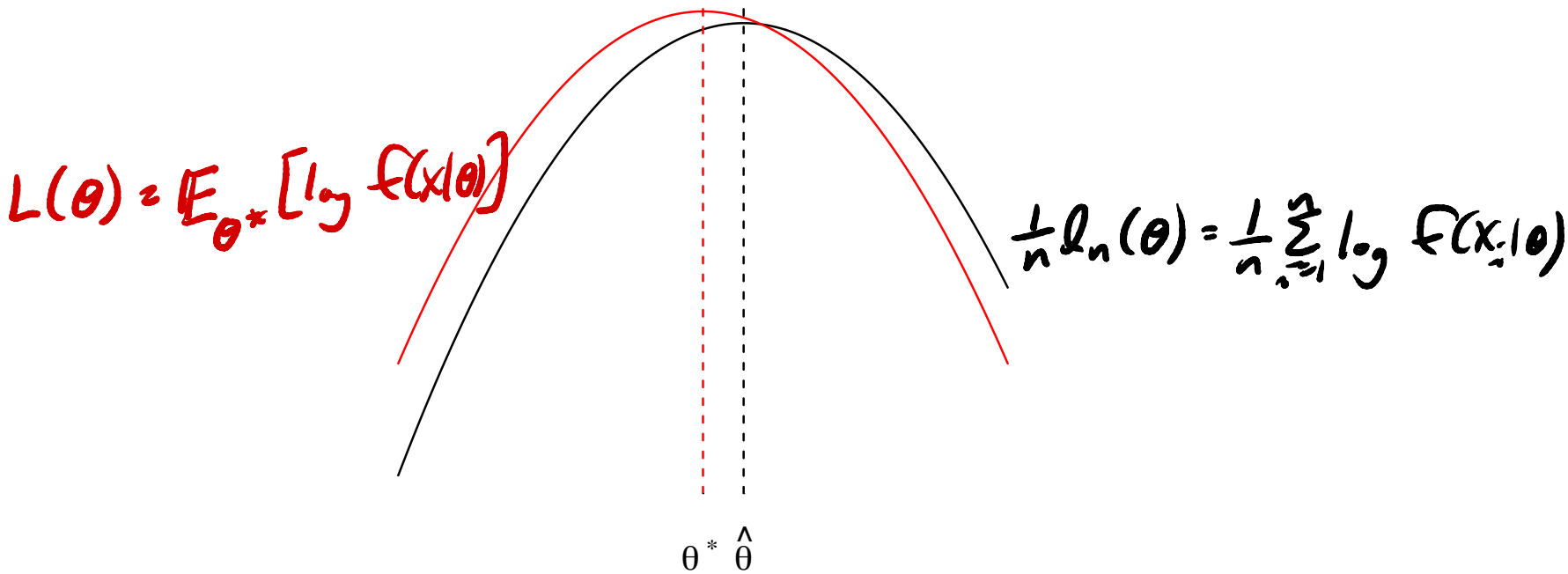
in distribution as $n \to \infty$, where $I(\theta)$ is the *Fisher information*, defined by

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

Thus the distribution of $\hat{\theta}$ for large n is approximately $\mathcal{N}(\theta, \frac{1}{nI(\theta)})$, and an asymptotic $(1 - \alpha)$-confidence interval for θ is

$$\hat{\theta} \pm z^{(1-\alpha/2)} \sqrt{\frac{1}{nI(\hat{\theta})}}$$

Expected log-likelihood function

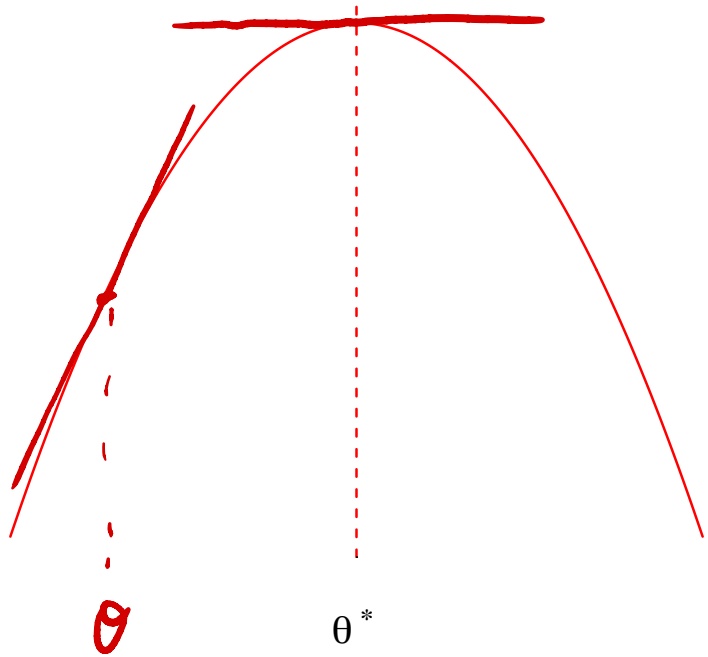


Recall that as $n \to \infty$, if the true parameter is θ^*, then the average log-likelihood approaches the *expected* log-likelihood

$$\frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta) \to L(\theta) = \mathbb{E}_{\theta^*}[\log f(X \mid \theta)]$$

in probability as $n \to \infty$. The expected log-likelihood $L(\theta)$ is maximized at the true parameter θ^*.

Expected score

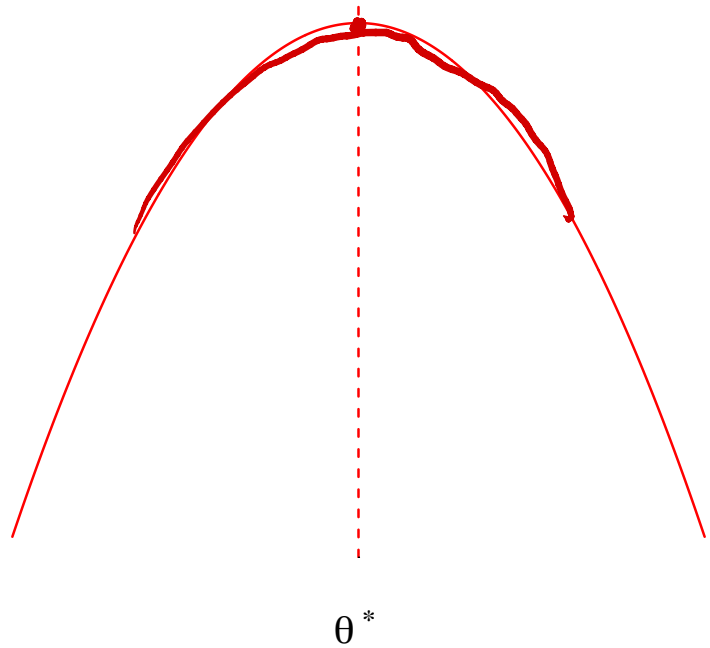


The derivative of $L(\theta)$ is

$$L'(\theta) = \frac{\partial}{\partial\theta}\mathbb{E}_{\theta^*}[\log f(X \mid \theta)] = \mathbb{E}_{\theta^*}[\underbrace{\frac{\partial}{\partial\theta}\log f(X \mid \theta)}_{=\text{score}}]$$

So $L'(\theta)$ is the expected score. Since $L(\theta)$ is maximized at θ^*, this also explains why the expected score is 0 at the true parameter θ^*.

Geometry of the Fisher information



The second derivative of $L(\theta)$ is

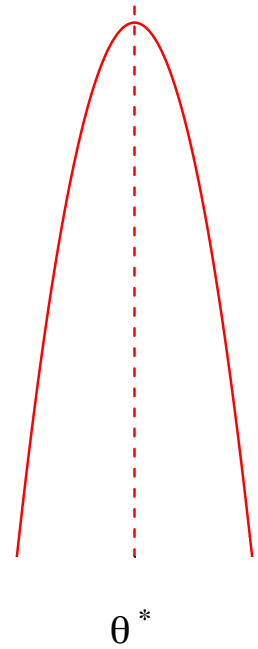
$$L''(\theta) = \frac{\partial^2}{\partial\theta^2}\mathbb{E}_{\theta^*}[\log f(X \mid \theta)] = \mathbb{E}_{\theta^*}[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)]$$

Then in particular, at the true parameter θ^*,

$$-L''(\theta^*) = -\mathbb{E}_{\theta^*}\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\Big|_{\theta=\theta_*}\right] = I(\theta^*)$$

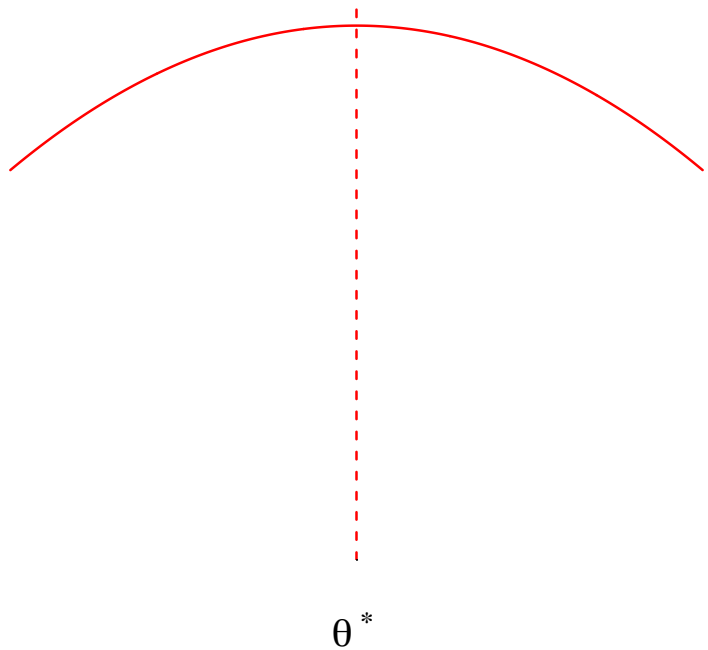
So $I(\theta^*)$ measures the curvature of $L(\theta)$ around its maximizer θ^*.

Example of large Fisher information



When Fisher information $I(\theta^*) = -L''(\theta^*)$ is large, moving θ slightly away from θ^* leads to a large decrease in the log-likelihood $L(\theta)$, i.e. a small change in θ makes the observed data much less probable. In this sense, the data contains more "information" that the parameter must be close to the true parameter θ^*.

Example of small Fisher information



Conversely, when Fisher information $I(\theta^*) = -L''(\theta^*)$ is small, large changes in θ only decrease the log-likelihood value $L(\theta)$ by a small amount, i.e. the observed data remains highly probable over large ranges of θ. In this sense, our data contains less "information" that the parameter is close to the true parameter θ^*.

Fisher information as an intrinsic property of the model

To summarize:

- $I(\theta)$ represents the curvature of the expected log-likelihood function around its maximum, when θ is the true parameter.
- $I(\theta)$ quantifies the rate at which the expected log-likelihood decreases in value around the true parameter, and thus may be interpreted as a measure of the expected amount of information the data contains about θ.
- We have shown that $I(\theta)$ is related to the asymptotic variance of the MLE, but it may also be understood as a fundamental quantity about the model $f(x \mid \theta)$ irrespective of any specific estimation procedure.

Cramer-Rao lower bound

Theorem (Cramer-Rao bound)

Consider a parametric model $f(x \mid \theta)$ where $\theta \in \mathbb{R}$ is a single parameter. Let $T(X_1, \ldots, X_n)$ be any unbiased estimator of θ based on data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$. Then

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

This is an example of an information-theoretic lower bound, stating that $\frac{1}{nI(\theta)}$ is a fundamental limit to how accurate *any* unbiased estimator of θ can be.[1]

[1] Related results show that no estimator — unbiased or not — can achieve a lower asymptotic mean-squared-error than $\frac{1}{nI(\theta)}$, except possibly at a measure-zero set of special values for θ. We will not discuss these results in our course.

Cramer-Rao lower bound

Proof: Define

$$Z = \frac{\partial}{\partial\theta} \ell_n(\theta) = \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial\theta} \log f(X_i|\theta)}_{\text{score at } X_i}$$

Recall, at the true parameter θ,

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right] = 0, \quad \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right] = I(\theta).$$

$$\Rightarrow \mathbb{E}_\theta[Z] = 0, \quad \mathrm{Var}_\theta[Z] = n \cdot I(\theta).$$

Correlation between Z and any unbiased estimator T is

$$\frac{\mathrm{Cov}_\theta[Z, T]}{\sqrt{\mathrm{Var}_\theta[Z]} \cdot \sqrt{\mathrm{Var}_\theta[T]}} \in [-1, 1]$$

Cramer-Rao lower bound

$$\Rightarrow \frac{\mathrm{Cov}_\theta[Z,T]^2}{\mathrm{Var}_\theta[Z]\cdot \mathrm{Var}_\theta[T]} \leq 1$$

$$\Rightarrow \mathrm{Var}_\theta[T] \geq \frac{\mathrm{Cov}_\theta[Z,T]^2}{\mathrm{Var}_\theta[Z]} = \frac{\mathrm{Cov}_\theta[Z,T]^2}{n\cdot I(\theta)}$$

T is an unbiased estimator of θ: For all θ,

$$\theta = \mathbb{E}_\theta[T] = \int T(x_1,\dots,x_n)\cdot f(x_1,\dots,x_n|\theta)\,dx_1\dots dx_n$$

Differentiate in θ:

$$1 = \int T(x_1,\dots,x_n)\cdot \underbrace{\frac{\partial}{\partial\theta} f(x_1,\dots,x_n|\theta)}_{= \frac{\partial}{\partial\theta}\log f(x_1,\dots,x_n|\theta)\,\cdot\, f(x_1,\dots,x_n|\theta)} dx_1\dots dx_n$$

Cramer-Rao lower bound

$$= \mathbb{E}_\theta\left[T(X_1,\dots,X_n)\cdot \underbrace{\frac{\partial}{\partial\theta}\log f(X_1,\dots,X_n|\theta)}_{=Z}\right]$$

$$\Rightarrow 1 = \mathbb{E}_\theta[TZ]$$

$$\text{So } \mathrm{Cov}_\theta[T,Z] = \mathbb{E}_\theta[TZ] - \mathbb{E}_\theta[T]\cdot \underbrace{\mathbb{E}_\theta[Z]}_{=0}$$

$$= \mathbb{E}_\theta[TZ] = 1.$$

$$\Rightarrow \mathrm{Var}_\theta[T] \geq \frac{1}{n\,I(\theta)}.$$

Asymptotic efficiency

An estimator $\hat{\theta}$ is **asymptotically efficient** if, in distribution as $n \to \infty$,

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

I.e., its variance for large n is approximately equal to the Cramer-Rao lower bound of $\frac{1}{nI(\theta)}$. Thus, our main theorem from last lecture shows that the MLE is asymptotically efficient.

If two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ based on $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ satisfy

$$\sqrt{n}(\hat{\theta}_1 - \theta) \to \mathcal{N}(0, v_1(\theta))$$
$$\sqrt{n}(\hat{\theta}_2 - \theta) \to \mathcal{N}(0, v_2(\theta))$$

as $n \to \infty$ where $v_2(\theta) < v_1(\theta)$, then $v_2(\theta)/v_1(\theta)$ is the **asymptotic relative efficiency** of $\hat{\theta}_1$ relative to $\hat{\theta}_2$.

Asymptotic relative efficiency

Since $\text{Var}[\hat{\theta}_1] \approx \frac{v_1(\theta)}{n}$ and $\text{Var}[\hat{\theta}_2] \approx \frac{v_2(\theta)}{n}$ for large n, relative efficiency may also be interpreted as the ratio of sample sizes needed for $\hat{\theta}_1$ and $\hat{\theta}_2$ to achieve the same variance: E.g. if $v_2(\theta) = 80\% \times v_1(\theta)$, then $\hat{\theta}_1$ using n samples has roughly the same variance as $\hat{\theta}_2$ using $80\% \times n$ samples.

Example: For $X_1, \ldots, X_n \overset{IID}{\sim} \text{Pareto}(\theta, 1)$, we have seen that the method-of-moments estimator $\hat{\theta}_{\text{MoM}}$ and MLE $\hat{\theta}_{\text{MLE}}$ satisfy

$$\sqrt{n}(\hat{\theta}_{\text{MoM}} - \theta) \to \mathcal{N}(0, \tfrac{\theta(\theta-1)^2}{\theta-2})$$
$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \to \mathcal{N}(0, \theta^2)$$

so the asymptotic relative efficiency of method-of-moments to maximum likelihood is $\frac{\theta(\theta-2)}{(\theta-1)^2}$ (which is less than 1).

Cramer-Rao bound for plug-in estimators

The Cramer-Rao lower bound holds also for plug-in estimators: Letting $\hat{\theta}$ be the MLE, the estimate $g(\hat{\theta})$ for $g(\theta)$ satisfies

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \to \mathcal{N}\left(0, \frac{g'(\theta)^2}{I(\theta)}\right)$$

by the delta method. Thus $\text{Var}[g(\hat{\theta})] \approx \frac{g'(\theta)^2}{nI(\theta)}$ for large n.

Theorem

For a parametric model $f(x \mid \theta)$ with a single parameter $\theta \in \mathbb{R}$, if $T(X_1, \ldots, X_n)$ is any unbiased estimator of $g(\theta)$ based on data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, then

$$\text{Var}_\theta[T] \geq \frac{g'(\theta)^2}{nI(\theta)}.$$

Efficiency of plug-in estimators

An estimator T for $g(\theta)$ based on $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ is asymptotically efficient if, in distribution as $n \to \infty$,

$$\sqrt{n}(T - g(\theta)) \to \mathcal{N}\left(0, \frac{g'(\theta)^2}{I(\theta)}\right)$$

Thus, the plug-in MLE $g(\hat{\theta})$ is asymptotically efficient.

Example: For $X_1, \ldots, X_n \overset{IID}{\sim}$ Pareto$(\theta, 1)$, the mean is $g(\theta) = \frac{\theta}{\theta-1}$. We have seen that

$$\sqrt{n}(\bar{X} - g(\theta)) \to \mathcal{N}(0, \tfrac{\theta}{(\theta-1)^2(\theta-2)})$$
$$\sqrt{n}(g(\hat{\theta}_{\text{MLE}}) - g(\theta)) \to \mathcal{N}(0, \tfrac{\theta^2}{(\theta-1)^4})$$

so the asymptotic relative efficiency of $\bar{X}$ to the plug-in MLE $g(\hat{\theta}_{\text{MLE}})$ is also $\frac{\theta(\theta-2)}{(\theta-1)^2}$.

Fisher information for multiple parameters

For a parametric model with k parameters $\theta \in \mathbb{R}^k$, the **Fisher information matrix** $I(\theta) \in \mathbb{R}^{k \times k}$ is the matrix whose (i, j) entry is defined by the equivalent expressions

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

For $k = 1$, this "1×1 matrix" $I(\theta)$ is the same as our previous definition of the Fisher information for a single parameter.

Asymptotic normality of the MLE

The inverse of this Fisher information matrix describes the asymptotic covariance matrix of the MLE $\hat{\theta} \in \mathbb{R}^k$.

Theorem

Let $f(x \mid \theta)$ be a parametric model where $\theta \in \mathbb{R}^k$. Let $\hat{\theta} \in \mathbb{R}^k$ be the MLE based on $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$. Then, under regularity assumptions,

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

The right side is a k-dimensional multivariate normal distribution, whose covariance $I(\theta)^{-1}$ is the $k \times k$ matrix inverse of $I(\theta)$.

MLE in the Gamma model

Example: Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$. Recall that the MLEs $(\hat{\alpha}, \hat{\beta})$ do not have closed-form expressions and are typically computed numerically.

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

$$\Rightarrow \log f(x \mid \alpha, \beta) = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log x - \beta x$$

$$\frac{\partial}{\partial \alpha} \log f(x \mid \alpha, \beta) = \log \beta - \underbrace{\frac{\partial}{\partial \alpha} \log \Gamma(\alpha)}_{\equiv \psi(\alpha) \text{ the digamma function}} + \log x$$

$$\frac{\partial}{\partial \beta} \log f(x \mid \alpha, \beta) = \frac{\alpha}{\beta} - x$$

MLE in the Gamma model

$$\frac{\partial^2}{\partial \alpha^2} \log f(x|\alpha,\beta) = -\underbrace{\frac{\partial^2}{\partial \alpha^2} \log \Gamma(\alpha)}_{\equiv \psi_1(\alpha) \text{ trigamma function}}$$

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

$$\text{So } I(\alpha,\beta) = -\mathbb{E}_{\alpha,\beta}\begin{bmatrix} \frac{\partial^2}{\partial \alpha^2} \log f(X|\alpha,\beta) & \frac{\partial^2}{\partial \alpha \partial \beta} \log f(X|\alpha,\beta) \\ \frac{\partial^2}{\partial \alpha \partial \beta} \log f(X|\alpha,\beta) & \frac{\partial^2}{\partial \beta^2} \log f(X|\alpha,\beta) \end{bmatrix}$$

$$= \begin{bmatrix} \psi_1(\alpha) & -1/\beta \\ -1/\beta & \alpha/\beta^2 \end{bmatrix}$$

$$\Rightarrow I(\alpha,\beta)^{-1} = \frac{1}{\psi_1(\alpha)\alpha/\beta^2 - 1/\beta^2} \cdot \begin{bmatrix} \alpha/\beta^2 & 1/\beta \\ 1/\beta & \psi_1(\alpha) \end{bmatrix}$$

MLE in the Gamma model

Informally, for large n, the MLEs $\hat{\alpha}$ and $\hat{\beta}$ have an approximate bivariate normal distribution

$$\mathcal{N}\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix},\ \frac{1}{n} I(\alpha,\beta)^{-1}\right)$$

The approximate variance of $\hat{\alpha}$ is the upper-left entry of $\frac{1}{n} I(\alpha,\beta)^{-1}$, which is $\frac{1}{n} \cdot \frac{\alpha}{\psi_1(\alpha)\alpha - 1}$.

The approximate covariance of $\hat{\alpha}$ and $\hat{\beta}$ is the off-diagonal entry of $\frac{1}{n} I(\alpha,\beta)^{-1}$, which is $\frac{1}{n} \cdot \frac{\beta}{\psi_1(\alpha)\alpha - 1}$. This is always positive, implying that the errors $\hat{\alpha} - \alpha$ and $\hat{\beta} - \beta$ are positively correlated for large n.

Comparison of variances when a parameter is known

Consider any model with two parameters (α, β) and Fisher information matrix

$$I(\alpha, \beta) = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$$

In the joint MLE $(\hat{\alpha}, \hat{\beta})$, the asymptotic variance of $\hat{\alpha}$ is the upper-left entry of $\frac{1}{n} I(\alpha, \beta)^{-1}$, which is

$$\frac{1}{n} \cdot \frac{1}{a - c^2/b}$$

Suppose instead that β is known, and let $\hat{\alpha}'$ be the MLE in model with a single parameter $\alpha \in \mathbb{R}$. The Fisher information in this one-parameter model is the upper-left entry of $I(\alpha, \beta)$, so the asymptotic variance of $\hat{\alpha}'$ is

$$\frac{1}{n} \cdot \frac{1}{a}$$

Comparison of variances when a parameter is known

Here $a - c^2/b > 0$ and $a, b > 0$, so

$$\frac{1}{a - c^2/b} > \frac{1}{a}$$

always. The difference between $\frac{1}{a-c^2/b}$ and $\frac{1}{a}$ represents the difference in asymptotic variance of the MLE for estimating α in settings when β is unknown vs. when β is known.

This is an example of a trade-off between model complexity and accuracy of estimation: A complex model with more parameters might better capture the true distribution of data, but each individual parameter may be more difficult to estimate than in a simpler model with fewer parameters.