11
The Information Inequality

11.1 The Score Function and Information

Now, for a model $\mathbf{X} \sim p(\mathbf{X};\theta), \theta \in \Theta$ for the random vector $\mathbf{X}$, define the score function,

$$U(\mathbf{X},\theta) = \frac{\partial \log p(\mathbf{X};\theta)}{\partial \theta} = \left(\frac{\partial p(\mathbf{X},\theta)}{\partial \theta}\right) / p(\mathbf{X},\theta). \tag{11.1}$$

Note that this is a function over Θ, only one of which will correspond to the true parameter value. We illustrate the use of the score function when the model is a family of density functions; the same holds for a family of mass functions.

For example, for the normal $\mathcal{N}(\mu,\sigma^2)$ family, assuming for the moment that σ^2 is known, we have

$$p(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

so

$$\log p(x;\mu) = -\log\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2}(x-\mu)^2$$

and therefore

$$U(X;\mu) = \frac{1}{\sigma^2}(X-\mu).$$

Note that

$$E_\mu\left(\frac{1}{\sigma^2}(X-\mu)\right)=0,$$

that is, the score has mean zero, when evaluated at the parameter which generated the data.

This mean zero property holds under smoothness in general, as the following argument shows, as long as we can differentiate under the integral as follows.

$$\begin{aligned}
E_\theta U(\mathbf{X},\theta) &= \int U(\mathbf{x},\theta)p(\mathbf{x},\theta)\\
&= \int \frac{\partial \log p(\mathbf{x};\theta)}{\partial\theta}p(\mathbf{x},\theta)\\
&= \int \frac{p'(\mathbf{x},\theta)}{p(\mathbf{x},\theta)}p(\mathbf{x},\theta)\\
&= \int p'(\mathbf{x},\theta)d\mathbf{x}\\
&= \frac{d}{d\theta}\int p(\mathbf{x},\theta)d\mathbf{x}\\
&= \frac{d}{d\theta}1=0.
\end{aligned}$$

Next, define the information as the variance of the score function as follows,

$$I(\theta)=\mathrm{Var}_\theta\left(U(\mathbf{X},\theta)\right).$$

As the score function has mean zero, its variance equals its second moment, and hence we may also write, using (11.1),

$$I(\theta) = E_\theta\left(U(\mathbf{X},\theta)\right)^2=\int\left(\frac{\partial p(\mathbf{x},\theta)}{\partial\theta}\right)^2/p(\mathbf{x},\theta)d\mathbf{x}. \tag{11.2}$$

If we have some additional smoothness, usually assumed, and which are known to hold for members of the exponential family, we have an alternative form of the information,

$$I(\theta)=-E_\theta\left(\frac{\partial^2\log p(\mathbf{X},\theta)}{\partial\theta^2}\right). \tag{11.3}$$

In particular, letting p' denote the partial of p with respect to θ,

$$\frac{\partial^2\log p(\mathbf{x},\theta)}{\partial\theta^2}=\frac{\partial}{\partial\theta}\frac{p'(\mathbf{x},\theta)}{p(\mathbf{x},\theta)}=\frac{p''(\mathbf{x},\theta)p(\mathbf{x},\theta)-p'(\mathbf{x},\theta)^2}{p(\mathbf{x},\theta)^2}.$$

Upon taking expectation, when multiplying by $p(\mathbf{x},\theta)$ the first term becomes simply the integral of $p''(\mathbf{x},\theta)$, and again by differentiating under

the integral as before, we have

$$0 = \int \frac{\partial^2 p(\mathbf{x}, \theta)}{\partial \theta^2}.$$

Note now that integrating the second term leads to the negative of (11.2), verifying (11.3).

Returning to the normal example, since the score function is

$$U(X, \mu) = \frac{1}{\sigma^2}(X - \mu)$$

the information is given by

$$I_X(\mu) = \text{Var}\left(\frac{1}{\sigma^2}(X - \mu)\right) = \frac{1}{\sigma^2}.$$

Note this expression makes good sense, since the smaller the variance, the more information there is in the single observation. We verify now that using the second derivative formula we get the same result. In particular, taking another derivative we obtain

$$\frac{\partial U(X, \mu)}{\partial \mu} = -\frac{1}{\sigma^2},$$

which, now taking an unnecessary expectation, is indeed the negative information.

Lets consider now some properties of the score. Let $X_1, \ldots, X_n$ be independent, not necessarily identically distributed, and $U_i(X_i, \theta)$ the score function for $p_i(x_i; \theta)$, the density of X_i. Then, since

$$p(\mathbf{X}, \theta) = \prod_{i=1}^{n} p_i(X_i; \theta) \quad \text{and hence} \quad \log p(\mathbf{X}, \theta) = \sum_{i=1}^{n} \log p_i(X_i; \theta),$$

taking partial with respect to θ we have

$$U(\mathbf{X}, \theta) = \sum_{i=1}^{n} U_i(X_i, \theta).$$

In other words, the score function for a collection of independent variables, whose distribution depends on a common θ, is the sum of the score functions for the independent variables.

As a consequence, the information is additive. Since the $X_1, \ldots, X_n$ independent imply that $U(X_1; \theta), \ldots, U(X_n; \theta)$ are independent, and the variance of the sum of independent variables is the sum of their variances,

we obtain

$$\begin{aligned} I_{\mathbf{X}}(\theta) &= \mathrm{Var}_\theta(U(\mathbf{X};\theta)) \\ &= \mathrm{Var}_\theta\left(\sum_{i=1}^n U(X_i;\theta)\right) \\ &= \sum_{i=1}^n \mathrm{Var}_\theta\left(U(X_i;\theta)\right) \\ &= \sum_{i=1}^n I_{X_i}(\theta). \end{aligned}$$

In particular, in the case where $X_1, \ldots, X_n$ are independent and identically distributed, we have

$$I_{\mathbf{X}}(\theta) = nI_X(\theta).$$

11.2 The Information Inequality

The information inequality, or Cramer Rao bound, is a consequence of the Cauchy Schwarz inequality which states that for random variables X and Y with finite second moment,

$$|EXY| \leq \sqrt{EX^2EY^2},$$

with equality if and only if there exist α and β, not both zero, and a constant c such that

$$P(\alpha X + \beta Y = c) = 1,$$

where c must necessarily be $E(\alpha X + \beta Y)$. Applying the inequality to $X - EX$ and $Y - EY$, we obtain

$$|\mathrm{Cov}(X,Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$$

with equality if and only if there exist α and β, not both zero, such that

$$P(\alpha(X - EX) + \beta(Y - EY) = 0) = 1.$$

For $\theta \in \Theta \subset \mathbb{R}$, that is, for a one dimensional parameter space, the Cramer Rao bound is as follows. First note that if X and Y are random variables with finite second moment, then generally

$$\mathrm{Cov}(X,Y) = E(X - EX)(Y - EY) = EXY - EXEY$$

so if $EY = 0$ we have

$$\mathrm{Cov}(X,Y) = EXY.$$

Now let $T(\mathbf{X})$ be a statistic with mean $g(\theta) = E_\theta T(\mathbf{X})$. Then differentiating

$$g(\theta) = \int T(\mathbf{x}) p(\mathbf{x};\theta) d\mathbf{x}$$

under the integral and using the fact that the score function has mean zero, we obtain

$$\begin{aligned}
|g'(\theta)| &= |\frac{d}{d\theta} \int T(\mathbf{x}) p(\mathbf{x};\theta) d\mathbf{x}| \\
&= |\int T(\mathbf{x}) \frac{\partial}{\partial\theta} p(\mathbf{x};\theta) d\mathbf{x}| \\
&= |\int T(\mathbf{x}) \frac{\frac{\partial}{\partial\theta} p(\mathbf{x};\theta)}{p(x;\theta)} p(x,\theta) d\mathbf{x}| \\
&= |E_\theta T(\mathbf{X}) U(\mathbf{X},\theta)| \\
&= |\mathrm{Cov}_\theta(T(\mathbf{X}), U(\mathbf{X},\theta))| \\
&\leq \sqrt{\mathrm{Var}_\theta(T(\mathbf{X}))\mathrm{Var}_\theta(U(\mathbf{X},\theta))} \\
&= \sqrt{\mathrm{Var}_\theta(T(\mathbf{X}))I_{\mathbf{X}}(\theta)},
\end{aligned}$$

Squaring and rearranging, we obtain the following lower bound on the variance of $T(\mathbf{X})$,

$$\mathrm{Var}_\theta(T(\mathbf{X}) \geq \frac{g'(\theta)^2}{I_{\mathbf{X}}(\theta)}.$$

Recall that we have equality if and only if $T(\mathbf{X})$ is linearly related to $U(\mathbf{X};\theta)$. Hence, we now have two additional ways to prove that an estimator is UMVU. We can now show that $T(\mathbf{X})$ achieves the lower bound, and also that $T(\mathbf{X})$ and $U(\mathbf{X};\theta)$ are linearly related. The condition for equality shows that there are not many UMVU's for a particular one dimensional model, as these all must be simply some multiple of the score function. In higher dimensions, there are more possibilities, however.

Consider the one parameter family, which is a special case of the beta, which, for $\alpha > 0$ is given by

$$p(x;\alpha) = \alpha x^{\alpha-1} \quad 0 < x < 1.$$

for which the score function is given by

$$U(x;\alpha) = \frac{\partial}{\partial\alpha} \log p(x;\alpha) = \log x + \frac{1}{\alpha}.$$

For the sample $X_1, \ldots, X_n$, the score function is sum of the marginal scores, so

$$U(\mathbf{X},\theta) = \sum_{i=1}^n \log X_i + \frac{n}{\alpha} = n\left(\frac{1}{n}\log\prod_{i=1}^n X_i + \frac{1}{\alpha}\right).$$

Hence any linear function of $\frac{1}{n}\log\prod_{i=1}^{n} X_i$ is UMVU of its expectation. Since the score has mean zero,

$$E_\alpha\left(-\frac{1}{n}\log\prod_{i=1}^{n} X_i\right) = \frac{1}{\alpha},$$

so

$$T(\mathbf{X}) = -\frac{1}{n}\log\prod_{i=1}^{n} X_i$$

is UMVU of $1/\alpha$.

Lets return to the normal example with known variance. The score for a single observation is

$$U(X;\mu) = \frac{1}{\sigma^2}(X-\mu), \tag{11.4}$$

so taking the sum,

$$U(\mathbf{X};\mu) = \frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j-\mu) = \frac{n}{\sigma^2}(\overline{X}-\mu).$$

Hence $U(\mathbf{X};\mu)$ and $\overline{X}$ are linearly related, so $\overline{X}$ must be UMVU for its expectation.

For the normal family where we take σ^2 are our unknown parameter, supposing for the moment that μ is known, recall

$$\log p(x;\theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2.$$

Taking partial with respect to σ^2 (not with respect to σ), we find that the score function for σ^2 is given by

$$U(X;\sigma^2) = \frac{\partial}{\partial\sigma^2}\log p(x;\theta) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2. \tag{11.5}$$

Noting that $E(X-\mu)^2 = \sigma^2$, we can check that the expectation of the score is zero, that is,

$$E_{\sigma^2}U(X;\sigma^2) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\sigma^2 = 0.$$

To compute the information, since the first term in the score (11.5) is constant, we have

$$I_X(\sigma^2) = \mathrm{Var}_{\sigma^2}(\frac{1}{2\sigma^4}(X-\mu)^2) = \frac{1}{4\sigma^4}\mathrm{Var}_{\sigma^2}(\frac{X-\mu}{\sigma})^2 = \frac{1}{2\sigma^4},$$

since $[(X-\mu)/\sigma]^2 \sim \chi^2_1$, and has variance

$$\mathrm{Var}(Z^2) = EZ^4 - (EZ)^2 = 3\cdot 1 - 1 = 2.$$

We can alternatively find the information by taking the expectation of the negative of the second partial. Differentiating (11.5) in σ^2, we obtain

$$\begin{aligned} -I_{\sigma^2}(X) &= E_{\sigma^2}\left(\frac{\partial^2}{\partial(\sigma^2)^2}\log p(x;\theta)\right) \\ &= E_{\sigma^2}\left(\frac{1}{2\sigma^4}-\frac{1}{\sigma^6}(x-\mu)^2\right) \\ &= \frac{1}{2\sigma^4}-\frac{\sigma^2}{\sigma^6}=-\frac{1}{2\sigma^4}, \end{aligned}$$

agreeing with the previous calculation.

When μ is known, then the estimator

$$T(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{11.6}$$

is UMVU for σ^2, since it is linearly related to the score function as follows. Summing (11.5) we obtain

$$U(\mathbf{X};\theta) = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2=-\frac{n}{2\sigma^2}+\frac{n}{2\sigma^4}T(\mathbf{X}).$$

Alternatively, we see that

$$\frac{n}{\sigma^2}T(\mathbf{X})=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2_n,$$

and therefore (with the 2 below the same 2 as the one above)

$$\mathrm{Var}_{\sigma^2}(\frac{n}{\sigma^2}T(\mathbf{X}))=2n$$

so that

$$\mathrm{Var}_{\sigma^2}(T(\mathbf{X}))=\frac{2\sigma^4}{n}=(nI_X(\sigma^2))^{-1}=(I_{\mathbf{X}}(\sigma^2))^{-1},$$

that is, $T(\mathbf{X})$ achieves the information bound. We remark that one also has that $T(\mathbf{X})$ is UMVU by the Lehmann Schefe theorem.

Now, what if μ and σ^2 are both unknown. On the one hand, we can still estimate μ by $\overline{X}$, here it seems not to matter whether we know σ^2 or not, we should do just as well, but we can no longer estimate σ^2 by $T(\mathbf{X})$ of (11.6). Actually, from the Lehmann Schefe theorem, we already know that

$$S^2=\frac{1}{n-1}\sum_{j=1}^{n}(X_i-\overline{X})^2$$

is UMVU in the case where μ is unknown, as it is an unbiased estimator which is function of a complete sufficient statistic. Note that when μ is known then these same statistics are not complete. At present we cannot

apply our information bound to the case where both parameters are unknown, but we may wonder if there exist cases where the lower bound is not achieved.

Consider in general the information for the location parameter in a location scale family. Taking $p(x), x \in \mathbb{R}$ any smooth density which is positive on $\mathbb{R}$, for $\theta \in \Theta \subset \mathbb{R}$ let $p(x;\theta) = (1/\sigma)p((x-\mu)/\sigma)$. The score function for μ is

$$\frac{\partial}{\partial \mu}\left(-\log\sigma + \log p((x-\mu)/\sigma)\right) = -\frac{(1/\sigma)p'((x-\mu)/\sigma)}{p((x-\mu)/\sigma)}$$

and the variance of this mean zero score is its second moment, we obtain its variance by squaring and multiplying by the density, which has an additional factor of $1/\sigma$, we have

$$I(\mu) = \frac{1}{\sigma^3}\int_{-\infty}^{\infty} \frac{[p'((x-\mu)/\sigma)]^2}{[p((x-\mu)/\sigma)]^2} p((x-\mu)/\sigma)dx$$
$$= \frac{1}{\sigma^3}\int_{-\infty}^{\infty} \frac{[p'((x-\mu)/\sigma)]^2}{p((x-\mu)/\sigma)}dx.$$

Making the change of variable $y = (x-\mu)/\sigma$, we have

$$I(\theta) = \frac{1}{\sigma^2}\int_{-\infty}^{\infty} \frac{[p'(y)]^2}{p(y)}dy.$$

For the normal,

$$p(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$$

so

$$\int_{-\infty}^{\infty} \frac{[p'(y)]^2}{p(y)} = \int_{-\infty}^{\infty} \frac{\frac{y^2}{2\pi}e^{-y^2}}{\frac{1}{\sqrt{2\pi}}e^{-y^2/2}} = \int_{-\infty}^{\infty} \frac{y^2}{\sqrt{2\pi}}e^{-y^2/2} = 1,$$

recovering the earlier result that the information about the mean in one observation of a normal is $I(\mu) = 1/\sigma^2$.

Now we compute the information for the Cauchy, based on the density

$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2} \quad p'(x) = -\frac{1}{\pi}\frac{2x}{(1+x^2)^2}$$

so that the ratio

$$\frac{[p']^2}{p} = \frac{4}{\pi}\frac{x^2}{(1+x^2)^3}.$$

Nice calculus exercise, or better yet, by complex contour integral, or better yet using a symbolic package, we have

$$\frac{4}{\pi}\int_{-\infty}^{\infty} \frac{x^2}{(1+x^2)^3}dx = \frac{4}{\pi}\times\frac{\pi}{8} = \frac{1}{2},$$

yielding

$$I(\mu) = \frac{\sigma^2}{2}$$

for the Cauchy family, half of the information about the location parameter than for a normal family.

Note that under the smoothness conditions the lower bound for the variance of our estimates in the iid model is $1/(nI_X(\theta)$, so they can decay no faster than $1/n$, or, the standard deviation can decay no faster than $n^{-1/2}$. Before taking on the multiparameter information inequality, lets consider a case where the smoothness conditions are violated; it may be that we do better than rate $1/n$ given by the lower bound. Consider the iid $\mathcal{U}[0,\theta]$ model, where we know, from Lehmann Schefe theorem, that

$$T(\mathbf{X}) = \left(\frac{n+1}{n}\right) X_{(n)},$$

is UMVU for θ, as it is unbiased for θ, and is a function of the complete sufficient statistic $X_{(n)}$, the maximum order statistic. The statistic $T(\mathbf{X})$ has variance

$$\mathrm{Var}_\theta(T(\mathbf{X})) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}_\theta(X_{(n)}),$$

Using the unbiasedness of $T(\mathbf{X})$ we have

$$E_\theta X_{(n)} = \left(\frac{n}{n+1}\right)\theta,$$

and now recalling the density of the maximum order statistic,

$$E_\theta X_{(n)}^2 = \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1} dx = \frac{n}{\theta^n}\frac{\theta^{n+2}}{n+2} = \frac{n\theta^2}{n+2}.$$

Hence

$$\begin{aligned}
\mathrm{Var}_\theta(X_{(n)}) &= \frac{n\theta^2}{n+2} - \left(\frac{n}{n+1}\right)^2\theta^2 \\
&= \theta^2 n\left(\frac{(n+1)^2 - n(n+2)}{(n+2)(n+1)^2}\right) \\
&= \theta^2\frac{n}{(n+2)(n+1)^2}.
\end{aligned}$$

Hence

$$\mathrm{Var}(T(\mathbf{X})) = \theta^2\frac{(n+1)^2}{n^2}\left(\frac{n}{(n+2)(n+1)^2}\right) = \frac{\theta^2}{n(n+2)} \leq \frac{\theta^2}{n^2},$$

which has rate $O(n^{-2})$, or, in terms of the standard deviation,

$$\sqrt{\mathrm{Var}(T(\mathbf{X})} \leq \frac{\theta}{n},$$

much faster than in cases where smoothness implies the information bounds holds; compare to $\overline{X}$ in the normal model, at rate $1/\sqrt{n}$.

We now consider the multiple parameter Information Inequality. Parallel to the one dimensional case, in the multidimensional case we $\mathbf{X} \sim p(\mathbf{x}, \theta), \theta \in \mathbf{R}^d$. Again we consider the score function

$$U(\theta, \mathbf{X}) = \frac{\partial \log p(\mathbf{x}; \theta)}{d\theta} \in \mathbf{R}^{d \times 1},$$

the gradient with respect to θ, and define $\mathbf{I}(\theta) \in \mathbf{R}^{d \times d}$ the information matrix, which is its variance

$$\mathbf{I}(\theta) = \mathrm{Var}_\theta(U(\theta, \mathbf{X})).$$

Define the partial order on non-negative definite matrices. We say

$$\Sigma \geq \Gamma \quad \text{if} \quad \Sigma - \Gamma \quad \text{is positive definite,}$$

that is, if

$$\mathbf{a}^\mathsf{T} \Sigma \mathbf{a} \geq \mathbf{a}^\mathsf{T} \Gamma \mathbf{a} \quad \text{for all } \mathbf{a}.$$

Note that in this notation we have Σ is non-negative definite if and only if $\Sigma \geq 0$.

Say we want to estimate a vector

$$q(\theta) \in \mathbf{R}^{1 \times r} \quad \text{by the statistic} \quad T(\mathbf{X}) \in \mathbf{R}^{1 \times r}$$

with expectation given by

$$E_\theta T(\mathbf{X}) = g(\theta) = \int T(\mathbf{x}) p(\mathbf{x}; \theta) d\mathbf{x} = \int p(\mathbf{x}; \theta) T(\mathbf{x}) d\mathbf{x} \in \mathbf{R}^{1 \times r}.$$

For instance, for a $\Gamma(\alpha, \beta)$ we may want to estimate

$$q(\theta) = (\alpha\beta, \alpha\beta^2),$$

the mean and variance of the distribution.

The multiparameter information inequality says that, under regularity,

$$\mathrm{Var}_\theta(T(\mathbf{X})^\mathsf{T}) \geq \dot{g}(\theta)^\mathsf{T} \mathbf{I}(\theta)^{-1} \dot{g}(\theta).$$

Clearly this inequality reduces to the one already shown in the case $d = 1$. Note that the derivative

$$\dot{g}(\theta) \in \mathbf{R}^{d \times r},$$

so that each column of $\dot{g}$ is the gradient with respect to θ of the row entry of g.

For the proof, we assume we can differentiate under the integral, (and noting that the density $p(\mathbf{x}; \theta)$ is placed before $T(\mathbf{X})$ so that the dimensions

work out when we differentiate) we form

$$\begin{aligned}\dot{g}(\theta) &= \int \frac{\partial}{\partial\theta} p(\mathbf{x},\theta) T(\mathbf{X}) d\mathbf{x} \\ &= \int \frac{\frac{\partial}{\partial\theta} p(\mathbf{x},\theta)}{p(\mathbf{x};\theta)} T(\mathbf{X}) p(\mathbf{x};\theta) d\mathbf{x} \\ &= E_\theta U(\mathbf{X},\theta) T(\mathbf{X}) = \mathrm{Cov}(U(\mathbf{X},\theta), T^{\mathsf{T}}(\mathbf{X})).\end{aligned}$$

Now we use

$$\mathrm{Cov}(AX, BY) = A\mathrm{Cov}(X,Y)B^{\mathsf{T}} \quad \text{and} \quad \mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)^{\mathsf{T}}$$

to derive

$$\begin{aligned}& 0 \le \mathrm{Var}_\theta(T(\mathbf{X})^{\mathsf{T}} - \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}U(\theta,\mathbf{X}))) \\ = \;& \mathrm{Var}_\theta(T(\mathbf{X}))^{\mathsf{T}} - \mathrm{Cov}(T(\mathbf{X})^{\mathsf{T}}, \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}U(\theta,\mathbf{X})) \\ - \;& \mathrm{Cov}(\dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}U(\theta,\mathbf{X}), T(\mathbf{X})^{\mathsf{T}}) + \mathrm{Var}_\theta(\dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}U(\theta,\mathbf{X})) \\ = \;& \mathrm{Var}_\theta(T(\mathbf{X}))^{\mathsf{T}} - \mathrm{Cov}(T(\mathbf{X})^{\mathsf{T}}, U(\theta,\mathbf{X}))\mathbf{I}(\theta)^{-1}\dot{g}(\theta) \\ - \;& \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}\mathrm{Cov}(U(\theta,\mathbf{X}), T(\mathbf{X})^{\mathsf{T}}) + \mathrm{Var}_\theta(\dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}U(\theta,\mathbf{X})) \\ = \;& \mathrm{Var}_\theta(T(\mathbf{X}))^{\mathsf{T}} - \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}\dot{g}(\theta) \\ - \;& \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}\dot{g}(\theta) + \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}\mathrm{Var}_\theta(U(\theta,\mathbf{X}))\mathbf{I}(\theta)^{-1}\dot{g}(\theta) \\ = \;& \mathrm{Var}_\theta(T(\mathbf{X})^{\mathsf{T}}) - \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}(\theta)^{-1}\dot{g}(\theta),\end{aligned}$$

which is the desired result.

Lets consider the two dimensional case $d = 2$, when we interested in unbiased estimates of the components of the θ vector, that is, the values of the parameters themselves. Lets write

$$\mathbf{I} = \begin{bmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{bmatrix}$$

for the information matrix. If θ_2 is known, then only θ_1 is a parameter, and the previous one dimensional results apply to yield the bound

$$\mathrm{Var}_\theta(T(\mathbf{X})) \ge \frac{1}{I_{11}}.$$

When θ_2 is unknown, we derive a lower bound of the form

$$\mathrm{Var}_\theta(T(\mathbf{X})) \ge \frac{1}{I^*_{11}}.$$

for some number I^*_{11}, which we call the effective information. To find I^*_{11}, take $g(\theta) = \theta_1$, so that

$$\dot{g}(\theta) = (1,0)^{\mathsf{T}}$$

to obtain the lower bound

$$\mathrm{Var}_\theta(T(\mathbf{X})) \ge \dot{g}(\theta)^{\mathsf{T}}\mathbf{I}^{-1}\dot{g}(\theta) = (\mathbf{I}^{-1})_{11} = \frac{I_{22}}{I_{11}I_{22} - I_{12}^2}$$

as

$$\mathbf{I}^{-1} = \frac{1}{I_{11}I_{22} - I_{12}^2} \begin{bmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{bmatrix}.$$

Hence,

$$I_{11}^* = \frac{I_{11}I_{22} - I_{12}^2}{I_{22}} = I_{11} - \frac{I_{12}^2}{I_{22}} = I_{11} - I_{11}\frac{I_{12}^2}{I_{11}I_{22}} = I_{11}(1 - \rho^2),$$

where

$$\rho = \text{Cor}(U_1, U_2)$$

the correlation between the score for θ_1 and θ_2. Note that the presence of the unknown θ_2 decreases the available information, as clearly,

$$I_{11}^* \leq I_{11}.$$

Note that the amount that I_{11} gets cut down by due to the presence of the unknown parameters, which gets worse and worse as the scores get more and more correlated, until finally there is zero information, and the parameter is not identifiable.

How as it in the normal case that ignorance of σ^2 did not affect the estimation of μ, while ignore of μ did affect the estimation of σ^2. The first fact indicates that the information is diagonal, since the information for μ seemed to have not decreased when we consider σ^2 unknown, suggesting $I_{11}^* = I_{11}$, which is true only when $\mathbf{I}$ is diagonal. But the second fact, that ignorance of μ affects the estimation of σ^2, suggests just the opposite. Lets examine the score functions to see which is correct.

Recalling the scores for μ and for σ^2, from (11.4) and (11.5),

$$U(x;\mu) = \frac{1}{\sigma^2}(x - \mu),$$

and

$$U(x;\sigma^2) = \frac{\partial}{\partial\sigma^2}\log p(x;\theta) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2,$$

we have already determined the diagonal elements of I when considering the estimation of each parameter when the other is known, in particular, we found that

$$I_{11} = \frac{1}{\sigma^2} \quad \text{and} \quad I_{22} = \frac{1}{2\sigma^4}.$$

As for the diagonal element, note that the two score functions above (having mean zero) are uncorrelated, since

$$E(X - \mu)^k = 0 \quad \text{for all odd } k.$$

Hence the information matrix $\mathbf{I}$ is diagonal, in particular,

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \quad \text{and} \quad I^{-1}(\theta) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.$$

That $\mathbf{I}$ is diagonal is consistent with the fact that ignorance of σ^2 does not effect estimation of μ. But what about estimation of σ^2, with lower bound, whether μ is known or not, of

$$\mathrm{Var}(T(\mathbf{X})) \geq \frac{2\sigma^4}{n}.$$

We have already shown the bound is achieved when μ is known. When μ is unknown, the Lehmann Scheffe shows that

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is UMVU. Using our results we showed for quadratic forms on the multivariate normal distribution, we proved

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2_{n-1} \quad \text{so} \quad \mathrm{Var}_\theta\left(\frac{(n-1)}{\sigma^2}S^2\right) = 2(n-1).$$

$$\text{yielding} \quad \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}.$$

Since S^2 is UMVU, when μ is unknown the information bound is not achievable. Note the theorem in no way guarantees that there will exist an estimator which achieves the bound.

Though the bound is not achievable for finite samples, though as $n \to \infty$ the ratio $(n-1)/n \to 1$, so asymptotically it won't make much difference. Later we will example where where ignorance of one parameter makes a difference in the estimation of the other, even as $n \to \infty$.

Still with $d = 2$, lets consider an example where we are interested in some function $q(\theta)$ of the two parameters, rather than their values. Consider then the estimation of the *signal to noise ratio* from a normal sample, that is,

$$g(\mu, \sigma^2) = \frac{\mu}{\sigma};$$

the inverse of this quantity is known as the *coefficient of variation*. The gradient vector of g is

$$\dot{g}(\mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma} \\ \frac{-\mu}{2\sigma^3} \end{bmatrix}.$$

Multiplying, we find the bound of

$$\dot{g}(\mu, \sigma^2)^{\mathsf{T}}\mathbf{I}^{-1}\dot{g}(\mu, \sigma^2) = [\frac{1}{\sigma}, \frac{-\mu}{2\sigma^3}] \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} \\ \frac{-\mu}{2\sigma^3} \end{bmatrix} = 1 + \frac{1}{2}(\frac{\mu}{\sigma})^2.$$

Hence, for n iid observations, one has

$$\mathrm{Var}_\theta(\mathbf{X}) \geq \frac{1}{n}\left(1 + \frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2\right).$$

It is actually not clear whether one can find an unbiased estimator. Still, what if you use

$$T = \frac{\bar{X}}{S}.$$

Will be biased (by Jensen's inequality) but one can still find if it achieves the bound when compared to all estimators having the same expectation.

Lets consider the Gamma family with both parameters unknown,

$$p(x;\theta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \quad x > 0.$$

We introduce the digamma and trigamma functions, which are the the first and second derivative, respectively, of the log of gamma function

$$\psi(\alpha) = \frac{d\log\Gamma(\alpha)}{d\alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \quad \text{and} \quad \pounds(\alpha) = \psi'(\alpha).$$

To begin to compute the score functions for the Gamma family, we take log of the density and obtain

$$\log p(x;\theta) = (\alpha-1)\log x - x/\beta - \log\Gamma(\alpha) - \alpha\log\beta.$$

Taking partial with respect to α we obtain,

$$U_\alpha(x;\theta) = \frac{\partial}{\partial\alpha}\log p(x;\theta) = \log x - \psi(\alpha) - \log\beta,$$

and with respect to β,

$$U_\beta(x;\theta) = \frac{\partial}{\partial\beta}\log p(x;\theta) = \frac{x}{\beta^2} - \frac{\alpha}{\beta}. \tag{11.7}$$

Computing the information by the second derivative formula, we have

$$I_{\alpha\alpha} = -E_\theta\left(\frac{\partial^2}{\partial\alpha^2}\log p(X;\theta)\right) = \psi'(\alpha) = \pounds(\alpha).$$

Recalling the mean of the Gamma distribution is given by $E_\theta X = \alpha\beta$, or just using the fact that the score has mean zero, we have

$$I_{\beta\beta} = -E_\theta\left(\frac{\partial^2}{\partial\beta^2}\log p(X;\theta)\right) = -E_\theta\left(-\frac{2X}{\beta^3} + \frac{\alpha}{\beta^2}\right) = \frac{\alpha}{\beta^2}.$$

Lastly, the mixed partial derivative gives the off diagonal element, by, say, taking partial with respect to α in (11.7),

$$-E_\theta\left(\frac{\partial^2}{\partial\alpha\partial\beta}\log p(x;\theta)\right) = \frac{1}{\beta}.$$

In particular, then, the information matrix

$$\mathbf{I} = \begin{bmatrix} \pounds(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix}$$

is not diagonal. As the score functions are not linearly related, the matrix must be strictly positive definite, and in particular, its determinant is positive; this is quite hard to prove by simply using the definitions of the derivatives of the Gamma function.

Computing the inverse,

$$I(\theta)^{-1} = \begin{bmatrix} \mathcal{L}(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix}^{-1} = \frac{\beta^2}{\alpha\mathcal{L}(\alpha) - 1} \begin{bmatrix} \frac{\alpha}{\beta^2} & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \mathcal{L}(\alpha) \end{bmatrix}.$$

Note that the average of score functions $U_\beta(x, \theta)$ in (11.7) in β is linearly related to $\overline{X}$, having mean $\alpha\beta$. Hence,

$$T_\alpha(\mathbf{X}) = \frac{\overline{X}}{\alpha}$$

is UMVU of β when α is known. We may double check, by noting that

$$\mathrm{Var}(T_\alpha(\mathbf{X})) = \frac{\beta^2}{n\alpha} = (nI_{\beta\beta})^{-1},$$

that is, $T_\alpha(\mathbf{X})$ achieves the Cramer Rao bound for the estimation of β. But if α is unknown then we cannot form this estimator. This is in some sense similar to what happened in the normal case for the estimation of variance, but there the matrix was diagonal and we were able to achieve the bound asymptotically. In this case the off diagonal element of the information matrix is nonzero, and we get into some trouble with the estimation of β when α is unknown, even asymptotically.

In particular, the variance bound for β when α is not known is the corner entry $(I^{-1})_{22}$. Since $\mathbf{I}$ is positive definite, and the determinant of a positive definite matrix is positive, we have

$$\alpha\mathcal{L}(\alpha) - 1 > 0$$

and therefore

$$(I^{-1})_{22} = \frac{\beta^2\mathcal{L}(\alpha)}{n(\alpha\mathcal{L}(\alpha) - 1)} > \frac{\beta^2}{n\alpha},$$

that is, the best variance which can be achieved in this case is strictly larger than the variance which is achieved when α is known.

As an aside, regading the trigamma function, differentiating the Gamma function under the integral we obtain

$$\Gamma^{(m)}(\alpha) = \int_0^\infty [\log x]^m x^{\alpha-1} e^{-x} dx,$$

so that for $X \sim \Gamma(\alpha, 1)$

$$E[\log X]^m = \frac{\Gamma^{(m)}(\alpha)}{\Gamma(\alpha)} \quad \text{so in particular} \quad \mathcal{L}(\alpha) = \mathrm{Var}(X) \geq 0.$$

Regression example with normal errors, first lets consider σ^2 known,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{or that} \quad Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2),$$

how much does ignorance of β_0 hurt in the estimation of β_1. The density for one observation is given by

$$p(y_i, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2)$$

which has logarithm

$$\log p(y_i, \theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2,$$

and taking partial derivatives with respect to β_0 yields

$$\frac{\partial \log p}{\partial \beta_0} = \frac{1}{\sigma^2}(y - \beta_0 - \beta_1 x) \quad \frac{\partial^2 \log p}{\partial \beta_0^2} = -\frac{1}{\sigma^2}$$

and with respect to β_1,

$$\frac{\partial \log p}{\partial \beta_1} = \frac{1}{\sigma^2}(y - \beta_0 - \beta_1 x)x \quad \frac{\partial^2 \log p}{\partial \beta_1^2} = -\frac{x^2}{\sigma^2}$$

and a mixed partial of

$$\frac{\partial^2 \log p}{\partial \beta_1 \beta_0} = -\frac{x}{\sigma^2}.$$

Note that none of these quantities depend on the random y.

Hence

$$I(\beta) = \frac{n}{\sigma^2}\begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \bar{x}^2 \end{bmatrix} \quad \text{and} \quad I^{-1}(\beta) = \frac{\sigma^2}{n(\bar{x}^2 - \bar{x}^2)}\begin{bmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

So the variance bounds for the slope, in the cases where the intercept is known, and not known, respectively are

$$\frac{\sigma^2}{n\bar{x}^2} \quad \text{and} \quad \frac{\sigma^2}{n(\bar{x}^2 - \bar{x}^2)},$$

the latter of which which is clearly larger. In particular, these are equal if and only if $\bar{x} = 0$.