Fisher Information & Efficiency

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 Introduction

Let $f(x \mid \theta)$ be the pdf of X for $\theta \in \Theta$; at times we will also consider a sample $\mathbf{x} = \{X_1, \cdots, X_n\}$ of size $n \in \mathbb{N}$ with pdf $f_n(\mathbf{x} \mid \theta) = \prod f(x_i \mid \theta)$. In these notes we'll consider how well we can estimate θ or, more generally, some function $g(\theta)$, by observing X or $\mathbf{x}$.

Let $\lambda(X \mid \theta) = \log f(X \mid \theta)$ be the natural logarithm of the pdf, and let $\lambda'(X \mid \theta)$, $\lambda''(X \mid \theta)$ be the first and second partial derivative with respect to θ (*not* X). In these notes we will only consider "regular" distributions, those with continuously differentiable (log) pdfs whose support doesn't depend on θ and which attain a maximum in the interior of Θ (not at an edge) where $\lambda'(X \mid \theta)$ vanishes. That will include most of the distributions we consider (normal, gamma, poisson, binomial, exponential, ...) but not the uniform.

The quantity $\lambda'(X \mid \theta)$ is called the "Score" (sometimes it's called the "Score statistic" but really that's a misnomer, since it usually depends on the parameter θ and *statistics* aren't allowed to do that). For a random sample $\mathbf{x}$ of size n, since the logarithm of a product is the sum of the logarithms, the Score is the sum $\lambda_n'(\mathbf{x} \mid \theta) = (\partial/\partial\theta) \log f_n(\mathbf{x} \mid \theta) = \sum \lambda(X_n \mid \theta)$.

Usually the MLE $\hat\theta$ is found by solving the equation $\lambda_n'(\mathbf{x} \mid \hat\theta) = 0$ for the Score to vanish, but today we'll use it for other things. For fixed θ, and evaluated at the random variable X (or vector $\mathbf{x}$), the quantity $Z := \lambda'(X \mid \theta)$ (or $Z_n := \lambda_n'(\mathbf{x} \mid \theta)$) is a random variable; let's find its mean and variance. First consider a single observation, or $n = 1$.

Let's consider continuously-distributed random variable X (for discrete distributions, just replace the integrals below with sums). Since

$$1 = \int f(x \mid \theta)\, dx$$

for any pdf and for every $\theta \in \Theta$, we can take a derivative to find

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int f(x \mid \theta)\, dx \\ &= \int \frac{\frac{\partial}{\partial\theta} f(x \mid \theta)}{f(x \mid \theta)}\ f(x \mid \theta)\, dx \\ &= \int \lambda'(x \mid \theta)\ f(x \mid \theta)\, dx = \mathsf{E}_\theta \lambda'(X \mid \theta), \end{aligned} \tag{1}$$

so the Score always has mean zero. The same reasoning shows that, for random samples, $\mathsf{E}_\theta \lambda_n'(\mathbf{x} \mid \theta) = 0$. The variance of the Score is denoted

$$I(\theta) = \mathsf{E}_\theta\big[\lambda'(X \mid \theta)^2\big] \tag{2}$$

and is called the *Fisher Information function*. Differentiating (1) (using the product rule) gives us another way to compute it:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta} \int \lambda'(x \mid \theta)\ f(x \mid \theta)\, dx \\
&= \int \lambda''(x \mid \theta)\ f(x \mid \theta)\, dx + \int \lambda'(x \mid \theta)\ f'(x \mid \theta)\, dx \\
&= \int \lambda''(x \mid \theta)\ f(x \mid \theta)\, dx + \int \lambda'(x \mid \theta)\ \frac{f'(x \mid \theta)}{f(x \mid \theta)}\ f(x \mid \theta)\, dx \\
&= \mathsf{E}_\theta\left[\lambda''(X \mid \theta)\right] + \mathsf{E}_\theta\left[\lambda'(X \mid \theta)^2\right] \\
&= \mathsf{E}_\theta\left[\lambda''(X \mid \theta)\right] + I(\theta) \qquad \text{by (2), so} \\
I(\theta) &= \mathsf{E}_\theta\big[-\lambda''(X \mid \theta)\big].
\end{aligned} \tag{3}$$

Since $\lambda_n(\mathbf{x} \mid \theta) = \sum \lambda(X_i \mid \theta)$ is the sum of n iid RVs, the variance $I_n(\theta) = \mathsf{E}_\theta[\lambda_n'(\mathbf{x} \mid \theta)^2] = n\, I(\theta)$ for a random sample of size n is just n times the Fisher Information for a single observation.

1.1 Examples

Normal: For the $\mathsf{No}(\theta, \sigma^2)$ distribution with fixed $\sigma^2 > 0$,

$$\begin{aligned}
\lambda(x \mid \theta) &= -\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\theta)^2 \\
\lambda'(x \mid \theta) &= \frac{1}{\sigma^2}(x-\theta) \\
\lambda''(x \mid \theta) &= -\frac{1}{\sigma^2} \\
I(\theta) &= \mathsf{E}_\theta\left[\frac{(X-\theta)^2}{\sigma^4}\right] = \frac{1}{\sigma^2},
\end{aligned}$$

the "precision" (inverse variance). The same holds for a random sample $I_n(\theta) = n/\sigma^2$ of size n. Thus the variance (σ^2/n) of $\hat\theta_n = \bar{x}_n$ is exactly $1/I_n(\theta)$.

Poisson: For the $\mathsf{Po}(\theta)$,

$$\begin{aligned}
\lambda(x \mid \theta) &= x\log\theta - \log x! - \theta \\
\lambda'(x \mid \theta) &= \frac{x-\theta}{\theta} \\
\lambda''(x \mid \theta) &= -\frac{x}{\theta^2} \\
I(\theta) &= \mathsf{E}_\theta\left[\frac{(X-\theta)^2}{\theta^2}\right] = \frac{1}{\theta},
\end{aligned}$$

so again $\hat\theta_n = \bar{x}_n$ has variance $1/I_n(\theta) = \theta/n$.

Bernoulli: For the Bernoulli distribution $\mathsf{Bi}(1, \theta)$,

$$\begin{aligned}
\lambda(x \mid \theta) &= x \log \theta + (1 - x) \log(1 - \theta) \\
\lambda'(x \mid \theta) &= \frac{x}{\theta} - \frac{1 - x}{1 - \theta} \\
\lambda''(x \mid \theta) &= -\frac{x}{\theta^2} - \frac{1 - x}{(1 - \theta)^2} \\
I(\theta) &= \mathsf{E}_\theta \left[\frac{X(1 - \theta)^2 + (1 - X)\theta^2}{\theta^2 (1 - \theta)^2} \right] = \frac{1}{\theta(1 - \theta)},
\end{aligned}$$

so again $\hat{\theta}_n = \bar{x}_n$ has variance $1/I_n(\theta) = \theta(1 - \theta)/n$.

Exponential: For $X \sim \mathsf{Ex}(\theta)$,

$$\begin{aligned}
\lambda(x \mid \theta) &= \log \theta - x\theta \\
\lambda'(x \mid \theta) &= \frac{1}{\theta} - x \\
\lambda''(x \mid \theta) &= -\frac{1}{\theta^2} \\
I(\theta) &= \mathsf{E}_\theta \left[\frac{(1 - X\theta)^2}{\theta^2} \right] = \frac{1}{\theta^2},
\end{aligned}$$

so $I_n(\theta) = n/\theta^2$. This time the variance of $\hat{\theta} = 1/\bar{x}_n$, $\frac{\theta^2 n^2}{(n-1)^2(n-2)}$, is bigger than $1/I_n(\theta) = \theta^2/n$ (you can compute this by noticing that $\bar{x}_n \sim \mathsf{Ga}(n, n\theta)$ has a Gamma distribution, and computing arbitrary moments $\mathsf{E}[Y^p] = \beta^{-p}\Gamma(\alpha + p)/\Gamma(\alpha)$ for any $Y \sim \mathsf{Ga}(\alpha, \beta)$, $p > -\alpha$).

It turns out that this lower bound *always* holds.

1.2 The Information Inequality

Let $T(X)$ be any statistic with finite variance, and denote its mean by

$$m(\theta) = \mathsf{E}_\theta T(X).$$

By the triangle inequality, the square of the covariance of any two random variables is never more than the product of their variances (this is just another way of saying the correlation is bounded by ± 1). We're going to apply that idea to the two random variables $T(X)$ and $\lambda'(X \mid \theta)$ (whose variance is $\mathsf{V}_\theta[\lambda'(X \mid \theta)] = I(\theta)$):

$$\begin{aligned}
\mathsf{V}_\theta[T(X)] \cdot \mathsf{V}_\theta[\lambda'(X \mid \theta] &\geq \left[\mathsf{E}_\theta \left\{ [T(X) - m(\theta)] \cdot [\lambda'(X \mid \theta) - 0] \right\} \right]^2 \\
&= \left[\int \left\{ [T(x)] \cdot [\lambda'(x \mid \theta)] \right\} f(x \mid \theta)\, dx \right]^2 \\
&= \left[\int T(x)\, f'(x \mid \theta)\, dx \right]^2 = \left[m'(\theta) \right]^2,
\end{aligned}$$

so

$$\mathsf{V}_\theta[T(X)] \geq \frac{[m'(\theta)]^2}{I(\theta)}$$

or, for samples of size n,

$$\mathsf{V}_\theta[T_n(\mathbf{x})] \geq \frac{[m'(\theta)]^2}{n\, I(\theta)}.$$

For vector parameters $\theta \in \Theta \subset \mathbb{R}^d$ the Fisher Information is a *matrix*

$$\begin{aligned} I(\theta) &= \mathsf{E}_\theta[\nabla\lambda(x \mid \theta)\ \nabla\lambda(x \mid \theta)^\intercal] \\ &= \mathsf{E}_\theta[-\nabla^2\lambda(x \mid \theta)] \end{aligned}$$

where $\nabla f(\theta)$ denotes the gradient of a real-valued function $f : \Theta \to \mathbb{R}$, a vector whose components are the partial derivatives $\partial f(\theta)/\partial\theta_i$; where $x^\intercal$ denotes the $1\times d$ transpose of the $d\times 1$ column vector x; and where $\nabla^2 f(\theta)$ denotes the *Hessian*, or $d \times d$ matrix of mixed second partial derivatives.
If $T(X)$ was intended as an estimator for some function $g(\theta)$, the MSE is

$$\begin{aligned} \mathsf{E}_\theta\left\{[T_n(\mathbf{x}) - g(\theta)]^2\right\} &= \mathsf{V}_\theta[T_n(\mathbf{x})] + [m(\theta) - g(\theta)]^2 \\ &\geq \frac{[m'(\theta)]^2}{n\, I(\theta)} + [m(\theta) - g(\theta)]^2 \\ &= \frac{[\beta'(\theta) + g'(\theta)]^2}{n\, I(\theta)} + \beta(\theta)^2 \end{aligned}$$

where $\beta(\theta) \equiv m(\theta) - g(\theta)$ denotes the bias. In particular, for estimating $g(\theta) = \theta$ itself,

$$\mathsf{E}_\theta\left\{[T_n(\mathbf{x}) - \theta]^2\right\} \geq \frac{[\beta'(\theta) + 1]^2}{n\, I(\theta)} + \beta(\theta)^2$$

and, for unbiased estimators, the MSE is always at least:

$$\mathsf{E}_\theta\left\{[T_n(\mathbf{x}) - \theta]^2\right\} \geq \frac{1}{n\, I(\theta)}.$$

These lower bounds were long thought to be discovered independently in the 1940s by statisticians Harold Crámer and Calyampudi Rao, and so this is often called the "Crámer-Rao Lower Inequality," but ever since Erich Lehmann brought to everybody's attention their earlier discovery by Maurice Fréchet in the 1870s they became called the "Information Inequality." We saw in examples that the bound is exactly met by the MLEs for the mean in normal and Poisson examples, but the inequality is strict for the MLE of the rate parameter in an exponential (or gamma) distribution.

It turns out there is a simple criterion for when the bound will be "sharp," *i.e.*, for when an estimator will exactly attain this lower bound. The bound arose from the inequality $\rho^2 \leq 1$ for the covariance ρ of $T(X)$ and $\lambda'(X \mid \theta)$; this *in*equality will be an *equality* precisely when $\rho = \pm 1$, *i.e.*, when $T(X)$ can be written as an affine function $T(X) = u(\theta)\lambda'(X \mid \theta) + v(\theta)$ of the Score. The

coefficients $u(\theta)$ and $v(\theta)$ can depend on θ, but not on X, but any θ-dependence has to cancel so that $T(X)$ won't depend on θ (because it was a *statistic*). In the Normal and Poisson examples the statistic T was X, which could indeed be written as an affine function of $\lambda'(X \mid \theta) = (X-\theta)/\sigma^2$ for the Normal or $\lambda'(X \mid \theta) = (X-\theta)/\theta$ for the Poisson, while in the Exponential case $1/X$ cannot be written as an affine function of $\lambda'(X \mid \theta) = (1/\theta - X)$.

1.2.1 Multivariate Case

A multivariate version of the Information Inequality exists as well. If $\Theta \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$, and if $T: \mathcal{X} \to \mathbb{R}^n$ is an n-dimensional statistic for some $n \in \mathbb{N}$ for data $X \sim f(x \mid \theta)$ taking values in a space $\mathcal{X}$ of arbitrary dimension, define the mean function $m: \mathbb{R}^k \to \mathbb{R}^n$ by $m(\theta) := \mathsf{E}_\theta T(X)$ and its $n \times k$ Jacobian matrix by

$$J_{ij} := \partial m_i(\theta)/\partial\theta_j.$$

Then the multivariate Information Inequality asserts that

$$\mathsf{Cov}_\theta[T(X)] \geq J\ I(\theta)^{-1} J^\intercal$$

where $I(\theta) := \mathsf{Cov}_\theta[\nabla_\theta \log f(X \mid \theta)]$ is the Fisher information matrix, where the notation "$A \geq B$" for $n \times n$ matrices A, B means that $[A - B]$ is positive semi-definite, and where $C^\intercal$ denotes the $k \times n$ transpose of an $n \times k$ matrix C. This gives lower bounds on the variance of $z'T(X)$ for all vectors $z \in \mathbb{R}^n$ and, in particular, lower bounds for the variance of components $T_i(X)$.

Examples

Normal Mean & Variance

If both the mean μ and precision $\tau = 1/\sigma^2$ are unknown for normal variates $X_i \overset{\text{iid}}{\sim} \mathsf{No}(\mu, 1/\tau)$, the Fisher Information for $\theta = (\mu, \tau)$ is

$$I(\theta) = -\mathsf{E}\begin{bmatrix} \frac{\partial^2}{\partial\mu^2}\ell & \frac{\partial^2}{\partial\mu\partial\tau}\ell \\ \frac{\partial^2}{\partial\tau\partial\mu}\ell & \frac{\partial^2}{\partial\tau^2}\ell \end{bmatrix} = -\mathsf{E}\begin{bmatrix} -\tau & (X-\mu) \\ (X-\mu) & -\tau^{-2}/2 \end{bmatrix} = \begin{bmatrix} \tau & 0 \\ 0 & \tau^{-2}/2 \end{bmatrix}$$

where $\ell(\theta) = \frac{1}{2}[\log(\tau/2\pi) - \tau(X-\mu)^2]$ is the log likelihood for a single observation.

Multivariate Normal Mean

If the mean vector $\mu \in \mathbb{R}^k$ is unknown but the covariance matrix $\Sigma_{ij} = \mathsf{E}(X_i - \mu_i)(X_j - \mu_j)$ is known for a multivariate normal RV $X \sim \mathsf{No}(\mu, \Sigma)$, the $(k \times k)$-Fisher Information matrix for μ is

$$I_{ij}(\mu) = -\mathsf{E}\frac{\partial^2}{\partial\mu_i\partial\mu_j}\left\{-\tfrac{1}{2}\log\left|2\pi\Sigma\right| - \tfrac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right\} = \left(\Sigma^{-1}\right)_{ij},$$

so (as in the one-dimensional case) the Fisher Information is just the precision (now a matrix).

Gamma

If both the shape α and rate λ are unknown for gamma variates $X_i \overset{\text{iid}}{\sim} \mathsf{Ga}(\alpha, \lambda)$, the Fisher Information for $\theta = (\alpha, \lambda)$ is

$$I(\theta) = -\mathsf{E}\begin{bmatrix} \frac{\partial^2}{\partial\mu^2}\ell & \frac{\partial^2}{\partial\mu\partial\tau}\ell \\ \frac{\partial^2}{\partial\tau\partial\mu}\ell & \frac{\partial^2}{\partial\tau^2}\ell \end{bmatrix} = -\mathsf{E}\begin{bmatrix} -\psi'(\alpha) & \lambda^{-1} \\ \lambda^{-1} & -\alpha\lambda^{-2} \end{bmatrix} = \begin{bmatrix} \psi'(\alpha) & -\lambda^{-1} \\ -\lambda^{-1} & \alpha\lambda^{-2} \end{bmatrix}$$

where $\ell(\theta) = \alpha \log \lambda + (\alpha - 1) \log X - \lambda X - \log \Gamma(\alpha)$ is the log likelihood for a single observation, and where $\psi'(z) := [\log \Gamma(z)]''$ is the "trigamma function" (Abramowitz and Stegun, 1964, 6.4), the derivative of the "digamma function" $\psi(z) := [\log \Gamma(z)]'$ (Abramowitz and Stegun, 1964, 6.3).

1.3 Regularity

The Information Inequality requires some regularity conditions for it to apply:

- The Fisher Information exists— equivalently, the log likelihood $\ell(\theta) := \log f(x \mid \theta)$ has partial derivatives with respect to θ for every $x \in \mathcal{X}$ and $\theta \in \Theta$, and they are in $L_2\big(\mathcal{X}, f(x \mid \theta)\big)$ for every $\theta \in \Theta$.
- Integration (wrt x) and differentiation (wrt θ) commute in the expression

$$\nabla_\theta \int_{\mathcal{T}} T(x)\ f(x \mid \theta)\, dx = \int_{\mathcal{T}} T(x)\ \nabla_\theta f(x \mid \theta)\, dx.$$

This latter condition will hold whenever the support $\{x :\ f(x \mid \theta) > 0\}$ doesn't depend on θ, and $\log f(x \mid \theta)$ has two continuous derivatives wrt θ everywhere.

These conditions both fail for the case of $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Un}(0, \theta)$, the uniform distribution on an interval $[0, \theta]$, and so does the conclusion of the Information Inequality. In this case the MLE is $\hat{\theta}_n = X_n^*(X) := \max\{X_i :\ 1 \le i \le n\}$, the sample maximum, whose mean squared error

$$\mathsf{E}_\theta |\hat{\theta}_n - \theta|^2 = \frac{2\theta^2}{(n+1)(n+2)}$$

tends to zero at rate n^{-2} as $n \to \infty$, while the Information Inequality bounds that rate below by a multiple of n^{-1} for problems satisfying the Regularity Conditions.

2 Efficiency

An estimator $\delta(X)$ of $g(\theta)$ is called *efficient* if it satisfies the Information Inequality exactly; otherwise its (absolute) efficiency is defined to be

$$\text{Eff}(\delta) = \frac{\frac{[\beta'(\theta) + g'(\theta)]^2}{I(\theta)} + \beta(\theta)^2}{\mathsf{E}_\theta \left\{[\delta(X) - g(\theta)]^2\right\}}$$

or, if the bias $\beta(\theta) \equiv [\mathsf{E}_\theta \delta(X) - g(\theta)]$ vanishes,

$$= \frac{[g'(\theta)]^2}{I(\theta) \mathsf{V}_\theta[\delta(X)]}.$$

It is *asymptotically efficient* if the efficiency for a sample of size n converges to one as $n \to \infty$. This happens for the MLE $\hat{\theta}_n = 1/\bar{x}_n$ for an Exponential distribution above, for example, whose bias is $\beta_n(\theta) = \mathsf{E}[\frac{1}{\bar{x}_n} - \theta] = \frac{\theta}{n-1}$ and whose absolute efficiency is

$$\text{Eff}(\hat{\theta}_n) = \frac{\frac{[1/(n-1)+1]^2}{n/\theta^2} + \frac{\theta^2}{(n-1)^2}}{\frac{\theta^2 n^2}{(n-1)^2(n-2)} + \frac{\theta^2}{(n-1)^2}} = \frac{(n+1)(n-2)}{(n+2)(n-1)}.$$

This increases to one as $n \to \infty$, so $\bar{\theta}_n$ is *asymptotically* efficient.

3 Asymptotic Relative Efficiency: ARE

If one estimator δ_1 of a quantity $g(\theta)$ has MSE $\mathsf{E}_\theta|\delta_1(\mathbf{x}) - g(\theta)|^2 \approx c_1/n$ for large n, while another δ_2 has MSE $\mathsf{E}_\theta|\delta_2(\mathbf{x}) - g(\theta)|^2 \approx c_2/n$ with $c_1 < c_2$, then the first will need a smaller sample-size $n_1 = (c_1/c_2)n_2$ to achieve the same MSE as the second would achieve with a sample of size n_2. The ratio (c_2/c_1) is called the *asymptotic relative efficiency* (or ARE) of δ_1 wrt δ_2. For example, if $c_2 = 2c_1$, then δ_1 needs $c_1/c_2 = 0.5$ times the sample size, and is $(c_2/c_1) = 2$ times more efficient than δ_2.

We won't go into it much in this course, but it's interesting to know that the for estimating the mean θ of the normal distribution $\mathsf{No}(\theta, \sigma^2)$ the sample mean $\hat{\theta} = \bar{x}_n$ achieves the Information Inequality lower bound of $\mathsf{E}_\theta|\bar{x}_n - \theta|^2 = \sigma^2/n = 1/I_n(\theta)$, while the sample median $M(\mathbf{x})$ has approximate[1] MSE $\mathsf{E}_\theta|M(\frac{\mathbf{x}-\theta}{\sigma}) - \theta|^2 \approx \pi\sigma^2/2n$, so the ARE of the median to the mean is

$$\frac{\sigma^2/n}{\pi\sigma^2/2(n+2)} \to \frac{2}{\pi} \approx 0.6366,$$

so the sample median will require a sample-size about $\pi/2 \approx 1.57$ times larger than the sample mean would for the same MSE. The median does offer a significant advantage over the sample mean, however— it is *robust* against model misspecification, *e.g.*, it is relatively unaffected by a few percent (or up to half) of errors or contamination in the data. Ask me if you'd like to know more about that.

4 Change of Variables for Fisher Information

The Fisher information function $\theta \rightsquigarrow I(\theta)$ depends on the particular way the model is parameterized. For example, the Bernoulli distribution can be parametrized in terms of the success probability p, the logistic $\eta = \log(\frac{p}{1-p})$, or in angular form with $\theta = \arcsin(\sqrt{p})$. The Fisher Information functions for these various choices will differ, but in a very specific way.

Consider a statistical model that can be parametrized in either of two ways,

$$X \sim f(x \mid \theta) = g(x \mid \eta), \quad \text{with} \quad \theta = \phi(\eta), \qquad \eta = \psi(\theta)$$

for one-dimensional parameter vectors θ and ψ, related by invertible differentiable 1:1 transformations. Then the Fisher Information functions I^θ and I^η in these two parameterizations are related

[1] You can show this by considering $\Phi\big(M(\frac{\mathbf{x}-\theta}{\sigma})\big)$, which is the median of n iid uniform random variables and so has exactly a $\mathsf{Be}(\alpha, \beta)$ distribution with $\alpha = \beta = \frac{n+1}{2}$ distribution for odd n, hence variance $1/4(n+2)$. Then use the "Delta method" to relate the variance of $M(\frac{\mathbf{x}-\theta}{\sigma})$ to that of $\Phi(M(\frac{\mathbf{x}-\theta}{\sigma}))$, $\mathsf{V}[\Phi(M(\frac{\mathbf{x}-\theta}{\sigma}))] \approx [\Phi'(0)]^2\mathsf{V}[M(\frac{\mathbf{x}-\theta}{\sigma})] = \frac{1}{2\pi}\mathsf{V}[M(\frac{\mathbf{x}-\theta}{\sigma})]$, so $\frac{1}{4(n+2)} \approx \frac{1}{2\pi}\mathsf{V}[M(\frac{\mathbf{x}-\theta}{\sigma})] = \frac{1}{2\pi\sigma^2}\mathsf{V}[M(\mathbf{x})]$, or $\mathsf{V}[M(\mathbf{x})] \approx \frac{\pi\sigma^2}{2(n+2)}$.

by

$$
\begin{aligned}
I^{\theta}(\theta) &= \mathsf{E}\left\{\left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right)^2\right\} \\
&= \mathsf{E}\left\{\left(\frac{\partial}{\partial\theta}\log g(x \mid \psi(\theta))\right)^2\right\} \\
&= \mathsf{E}\left\{\left(\frac{\partial}{\partial\eta}\log g(x \mid \eta)\frac{\partial\eta}{\partial\theta}\right)^2\right\} \\
&= \mathsf{E}\left\{\left(\frac{\partial}{\partial\eta}\log g(x \mid \eta)\right)^2\right\}\left(\frac{\partial\eta}{\partial\theta}\right)^2 \\
&= I^{\eta}(\eta)\big(J^{\eta}_{\theta}\big)^2, \qquad (4)
\end{aligned}
$$

the Fisher information $I^{\eta}(\eta)$ in the η paramaterization times the square of the Jacobian $J^{\eta}_{\theta} := \partial\eta/\partial\theta$ for changing variables.

For example, the Fisher Information for Bernoulli random variable in the usual success-probability parametrization was shown in Section (1.1) to be $I^{p}(p) = 1/p(1-p)$. In the logistic parametrization it would be

$$
I^{\eta}(\eta) = \frac{e^{\eta}}{(1+e^{\eta})^2} = I^{p}\big(\frac{e^{\eta}}{1+e^{\eta}}\big)(J^{p}_{\eta})^2,
$$

for Jacobian $J^{p}_{\eta} = \partial p/\partial\eta = e^{\eta}/(1+e^{\eta})^2$ and inverse transformation $p = e^{\eta}/(1+e^{\eta})$, while in the arcsin parameterization the Fisher Information

$$
I^{\theta}(\theta) = I^{p}\big(\sin^2(\theta)\big)(J^{p}_{\theta})^2 = 4,
$$

a constant, for Jacobian $J^{p}_{\theta} = \partial p/\partial\theta = 2\sin\theta\cos\theta$ and inverse transformation $p = \sin^2\theta$.

4.1 Jeffreys' Rule Prior

From (4) it follows that the unnormalized prior distributions

$$
\pi^{\theta}_{J}(\theta) = \big(I^{\theta}(\theta)\big)^{1/2} \quad \text{and} \quad \pi^{\eta}_{J}(\eta) = \big(I^{\eta}(\eta)\big)^{1/2}
$$

are related by

$$
\pi^{\theta}_{J}(\theta) = \pi^{\eta}_{J}(\eta)\ |\partial\eta/\partial\theta|,
$$

exactly the way they should be related for the change of variables $\eta \rightsquigarrow \theta$.

It was this invariance property that led Harold Jeffreys (1946) to propose $\pi_{J}(\theta)$ as a default "objective" choice of prior distribution. Much later José Bernardo (1979) showed that this is also the "Reference prior" that maximizes the entropic distance from the prior to the posterior, *i.e.*, the information to be learned from an experiment; see (Berger et al., 2009, 2015) for a more recent and broader view. In estimation problems with one-dimensional parameters a Bayesian analysis using $\pi_{J}(\theta) \propto I(\theta)^{1/2}$ is widely regarded as a suitable objective Bayesian approach.

4.2 Examples using Jeffreys' Rule Prior

Normal: For the $\mathsf{No}(\theta, \sigma^2)$ distribution with known σ^2, the Fisher Information is $I(\theta) = 1/\sigma^2$, a constant (for known σ^2), so $\pi_J(\theta) \propto 1$ is the improper uniform distribution on $\mathbb{R}$ and the posterior distribution for a sample of size n is

$$\pi_J(\theta \mid \mathbf{x}) \sim \mathsf{No}(\bar{X}_n, \sigma^2/n)$$

with posterior mean $\hat{\theta}_J = \bar{X}_n$, the same as the MLE.

Poisson: For the $\mathsf{Po}(\theta)$, the Fisher Information is $I(\theta) = 1/\theta$, so $\pi_J(\theta) \propto \theta^{-1/2}$ is the improper conjugate $\mathsf{Ga}(1/2, 0)$ distribution and the posterior for a sample $\mathbf{x}$ of size n is

$$\pi_J(\theta \mid \mathbf{x}) \sim \mathsf{Ga}\big(\tfrac{1}{2} + \sum X_i, n\big),$$

with posterior mean $\hat{\theta}_J = \bar{X}_n + 1/2n$, asymptotically the same as the MLE.

Bernoulli: For the $\mathsf{Bi}(1, \theta)$, the Fisher Information is $I(\theta) = 1/\theta(1-\theta)$, so $\pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ is the conjugate $\mathsf{Be}(1/2, 1/2)$ distribution and the posterior for a sample $\mathbf{x}$ of size n is

$$\pi_J(\theta \mid \mathbf{x}) \sim \mathsf{Ga}\big(\tfrac{1}{2} + \sum X_i, \tfrac{1}{2} + n - \sum X_i\big),$$

with posterior mean $\hat{\theta}_J = (\sum X_i + \frac{1}{2})/(n+1)$. In the logistic parametrization the Jeffreys' Rule prior is $\pi_J(\eta) \propto 1/(e^{\eta/2} + e^{-\eta/2}) \propto \mathrm{sech}(\eta/2)$, the hyperbolic secant, while the Jeffreys' Rule prior is uniform $\pi_J(\theta) \propto 1$ in the arcsin parametrization, but each of these induces the $\mathsf{Be}(\frac{1}{2}, \frac{1}{2})$ for p under a change of variables.

Exponential: For the $\mathsf{Ex}(\theta)$, the Fisher Information is $I(\theta) = 1/\theta^2$, so the Jeffreys' Rule prior is the scale-invariant improper $\pi_J(\theta) \propto 1/\theta$ on $\mathbb{R}_+$, with posterior density for a sample $\mathbf{x}$ of size n is

$$\pi_J(\theta \mid \mathbf{x}) \sim \mathsf{Ga}\big(n, \sum X_i\big),$$

with posterior mean $\bar{\theta}_J = 1/\bar{X}_n$ equal to the MLE.

4.3 Multidimensional Parameters

When θ and η are d-dimensional a similar change-of-variables expression holds, but now I^θ, I^η and the Jacobian J^η_θ are all $d \times d$ matrices:

$$\begin{aligned}
I^\theta_{ij}(\theta) &= \mathsf{E}\left\{\left(\frac{\partial}{\partial\theta_i}\log f(x \mid \theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(x \mid \theta)\right)\right\} \\
&= \mathsf{E}\left\{\left(\frac{\partial}{\partial\theta_i}\log g(x \mid \psi(\theta))\right)\left(\frac{\partial}{\partial\theta_j}\log g(x \mid \psi(\theta))\right)\right\} \\
&= \mathsf{E}\left\{\left(\sum_k \frac{\partial}{\partial\eta_k}\log g(x \mid \eta)\frac{\partial\eta_k}{\partial\theta_i}\right)\left(\sum_l \frac{\partial}{\partial\eta_l}\log g(x \mid \eta)\frac{\partial\eta_l}{\partial\theta_j}\right)\right\} \\
I^\theta(\theta) &= J^{\eta\intercal}_\theta I^\eta(\eta) J^\eta_\theta,
\end{aligned}$$

Jeffreys noted that now the prior distribution

$$\pi_J(\theta) \propto |\det I(\theta)|^{1/2}$$

proportional to the square root of the *determinant* of the Jacobian matrix is invariant under changes of variables, but both he and later authors noted that $\pi_J(\theta)$ has some undesirable features in $d \geq 2$ dimensions including the important example of $\mathsf{No}(\mu, \sigma^2)$ with unknown mean and variance. "Reference priors" (Berger and Bernardo, 1992a,b) are currently considered the best choice for objective Bayesian analysis in multi-parameter problems, but they are challenging to compute and to work with. The tensor product of independent one-dimensional Jeffreys priors $\pi_J(\theta_1)\pi_J(\theta_2)\cdots\pi_J(\theta_d)$ are frequently recommended as an acceptable alternative.

References

Abramowitz, M. and Stegun, I. A., eds. (1964), *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, *Applied Mathematics Series*, volume 55, Washington, D.C.: National Bureau of Standards, reprinted in paperback by Dover (1974); on-line at `http://www.math.sfu.ca/~cbm/aands/`.

Berger, J. O. and Bernardo, J. M. (1992a), "On the development of the Referrence Prior method," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 35–49.

Berger, J. O. and Bernardo, J. M. (1992b), "Ordered group reference priors with application to a multinomial problem," *Biometrika*, 79, 25–37.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009), "The Formal Definition of Reference Priors," *Annals of Statistics*, 37, 905–938.

Berger, J. O., Bernardo, J. M., and Sun, D. (2015), "Overall Objective Priors," *Bayesian Analysis*, 10, 189–221, doi:10.1214/14-BA915.

Bernardo, J. M. (1979), "Reference posterior distributions for Bayesian inference (with discussion)," *Journal of the Royal Statistical Society, Ser.* B*: Statistical Methodology*, 41, 113–147.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186, 453–461, doi:10.1098/rspa.1946.0056.