1. Fisher Information

Let $f(x \mid \theta)$ be a density function with the property that $\log f(x \mid \theta)$ is differentiable in θ throughout the open p-dimensional parameter set $\Theta \subset \mathbb{R}^p$; then the **score statistic** (or score function) is defined by

$$Z(X) \quad \equiv \quad \nabla_\theta \log f(x \mid \theta) = \quad \frac{\nabla_\theta f(x \mid \theta)}{f(x \mid \theta)}$$

and the **Fisher** (or **Expected**) **Information** matrix is defined by

$$I(\theta) \quad \equiv \quad \mathsf{E}\left[Z(X)Z(X)' \mid \theta\right];$$

if we may exchange integration with differentiation then we can calculate

$$\begin{aligned}
\mathsf{E}[Z_i(X) \mid \theta] &= \int_{\mathcal{X}} [\frac{d}{d\theta_i} \log f(x \mid \theta)]\, f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} \frac{\frac{d}{d\theta_i} f(x \mid \theta)}{f(x \mid \theta)}\, f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} \frac{d}{d\theta_i} f(x \mid \theta)\, dx \\
&= \frac{d}{d\theta_i} \int_{\mathcal{X}} f(x \mid \theta)\, dx \\
&= 0
\end{aligned}$$

and hence $\mathsf{E}[Z(X) \mid \theta] = 0$ and $\mathsf{Cov}[Z(X) \mid \theta] = \mathsf{E}\left[Z(X)Z(X)' \mid \theta\right] = I(\theta)$; taking another derivative with respect to θ_j of the equation $\mathsf{E}[Z_i(X) \mid \theta] = 0$ gives, by the product rule,

$$\begin{aligned}
0 &= \frac{d}{d\theta_j} \mathsf{E}[Z_i(X) \mid \theta] \\
&= \frac{d}{d\theta_j} \int_{\mathcal{X}} [\frac{d}{d\theta_i} \log f(x \mid \theta)]\, [f(x \mid \theta)]\, dx \\
&= \int_{\mathcal{X}} [\frac{d^2}{d\theta_i\, d\theta_j} \log f(x \mid \theta)]\, [f(x \mid \theta)]\, dx + \int_{\mathcal{X}} [\frac{d}{d\theta_i} \log f(x \mid \theta)]\, [\frac{d}{d\theta_j} \log f(x \mid \theta)]\, f(x \mid \theta)]\, dx \\
&= \mathsf{E}\left[\frac{d^2}{d\theta_i\, d\theta_j} \log f(x \mid \theta)\right] + I(\theta),
\end{aligned}$$

so we may also compute the Fisher Information as

$$I(\theta) \quad = \quad \mathsf{E}\left[-\nabla^2 \log f(x \mid \theta)\right]$$

as the matrix of expected negative second derivatives of the log likelihood with respect to θ.

The Fisher Information matrix depends on the parametrization chosen. If we rewrite our model in terms of some other parameter η, then by the vector chain rule $\nabla_\theta = J^\mathsf{T} \nabla_\eta$ (J denotes the Jacobian matrix $J = \frac{\partial \eta}{\partial \theta}$ with components $J_{ij} = \partial \eta_i / \partial \theta_j$), the Fisher Information for the two parametrizations are related by

$$I(\theta) = J^\mathsf{T}\, I(\eta)\, J. \tag{1}$$

1.1. The Information Inequality

Now let $\Theta \subset \mathbb{R}$ be one-dimensional and let T be any statistic with finite expectation $\psi(\theta) \equiv \mathsf{E}[T(X) \mid \theta]$, and assume additionally that ψ is differentiable throughout Θ to justify exchanging integration and differentiation as follows:

$$\begin{aligned}
\psi'(\theta) &= \frac{d}{d\theta} \int_{\mathcal{X}} T(x)\, f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} T(x)\, \frac{d}{d\theta} f(x \mid \theta)\, dx \\
&= \int_{\mathcal{X}} T(x)\, Z(x)\, f(x \mid \theta)\, dx \\
&= \mathsf{E}\left[T(X)\, Z(X) \mid \theta\right] = \mathsf{Cov}\left[T(X)\, Z(X)\right],
\end{aligned}$$

so the score statistic $Z(X) \equiv \frac{d}{d\theta} \log f(x \mid \theta)$ has mean zero, variance $I(\theta)$, and covariance $\psi'(\theta) = \mathsf{Cov}[T(X), Z(X)]$ with $T(X)$; by the Covariance Inequality $|\mathsf{Cov}(T, Z)|^2 \leq \mathsf{V}(T)\, \mathsf{V}(Z)$ (Minkowski's inequality), we can conclude that $|\psi'(\theta)|^2 \leq I(\theta) \mathsf{V}(T(X))$, or that

$$\mathsf{V}(T(X)) \geq \frac{|\psi'(\theta)|^2}{I(\theta)};$$

in particular, any unbiased estimator T of θ must have risk

$$R(T, \theta) \geq \frac{1}{I(\theta)}$$

bounded below by the celebrated Information Inequality. This result was commonly referred to as the Cramèr-Rao inequality, until Frechèt's earlier discovery was widely recognized.

2. Bayesian Central Limit Theorem

The *observed information* for a single observation $X = x$ from the model $X \sim f(x|\theta)$ is

$$i(\theta, x) \quad = \quad -\nabla_\theta^2 \log f(x \mid \theta);$$

evidently the Fisher (expected) information is related to this by $I(\theta) = \mathsf{E}[i(\theta, X)|\theta]$. The likelihood for a sample of size n is just the product of the individual likelihoods, leading to a *sum* for the *log* likelihoods, and observed information

$$i(\theta, x) \quad = \quad \sum_{j=1}^{n} i(\theta, x_j).$$

If the log likelihood $\log f(x|\theta)$ is differentiable throughout Θ and attains a unique maximum at an interior point $\hat{\theta}_n(x) \in \Theta$, then we can expand $\log f(x|\theta)$ in a second-order Taylor series for $\theta = \hat{\theta}_n(x) + \epsilon/\sqrt{n}$ close to $\hat{\theta}_n(x)$ to find

$$\begin{aligned}
\log f(x|\theta) &= \log f(x|\hat{\theta}_n) + \frac{(\epsilon/\sqrt{n})^1}{1!}\nabla_\theta \log f(x|\hat{\theta}_n) + \frac{(\epsilon/\sqrt{n})^2}{2!}\nabla_\theta^2 \log f(x|\hat{\theta}_n) \\
&\qquad + o\big((\epsilon/\sqrt{n})^2 |\nabla_\theta^2 \log f(x|\hat{\theta}_n)|\big) \\
&= \log f(x|\hat{\theta}_n) + 0 - \frac{\epsilon^2}{2}\frac{1}{n}\sum_{j=1}^{n} i(\hat{\theta}_n, x_j) + o(1) \\
&\to \log f(x|\hat{\theta}_n) - \frac{\epsilon^2}{2}\mathsf{E}[i(\hat{\theta}_n, X)] \\
&= \log f(x|\hat{\theta}_n) - \frac{\epsilon^2}{2} I(\theta),
\end{aligned}$$

where we have used the consistency of $\hat{\theta}_n$ and have applied the strong law of large numbers for $i(\theta, X)$. Thus we have the likelihood approximation $f(x|\theta) \approx \mathsf{No}(\hat{\theta}_n(x), nI(\hat{\theta}_n))$, normal with mean the MLE $\hat{\theta}_n(x)$ and *precision* $nI(\hat{\theta}_n)$ (or covariance $\frac{1}{n}I(\hat{\theta}_n)^{-1}$).

3. Exponential Families

Consider a sample $X = (X_1, \cdots, X_n)$ of some number n of independent replicates, all from the same probability distribution with pdf or pmf $f(x \mid \theta)$

of *exponential family* form

$$f(x \mid \theta) = h(x) \exp\left[\sum_{i=1}^{k} \eta_i(\theta) t_i(x) - c(\theta)\right];$$

since $1 \equiv \int_{\mathcal{X}} f(x \mid \theta)\, dx = e^{-c(\theta)} \int_{\mathcal{X}} h(x)\, e^{\eta(\theta)\cdot T(x)}\, dx$, $c(\theta)$ must be given by

$$c(\theta) \equiv \log\left(\int_{\mathcal{X}} h(x) e^{\eta(\theta)\cdot T(x)}\, dx\right).$$

Many of the commonly-considered distributions can be written as exponential families with $k = 1$ or 2, for suitable h, c, and $\{\eta_i, t_i\}_{i \le k}$. The likelihood function for a random sample of size n from the exponential family is

$$L(\theta) = \exp\left[\sum_{i=1}^{k} \eta_i(\theta) \sum_{j=1}^{n} t_i(x_j) - nc(\theta)\right],$$

which depends on the data only through the k-dimensional statistic T with components $T_i = \sum_{j \le n} t_i(x_j)$. This *natural sufficient statistic* (see below) summarizes the data completely for any inference about θ.

It is often convenient to reparametrize exponential families to the *natural parameter* $\eta = \eta(\theta) \in \mathbb{R}^k$, leading (with a slight abuse of notation for $c(\cdot)$) to

$$f(x \mid \eta) = h(x) e^{\eta\cdot T(x) - n\, c(\eta)}$$

for a sample of size n, where again $c(\eta) \equiv \log\left(\int_{\mathcal{X}} h(x) e^{\eta\cdot T(x)}\, dx\right)$. We can calculate the moment generating function (MGF) for $T(X)$ as

$$\begin{aligned}
M_T(s) &= \mathsf{E}\left[e^{s\cdot T(X)}\right] \\
&= \int_{\mathcal{X}} e^{s\cdot T(x) - n\, c(\eta)}\, h(x) e^{\eta\cdot T(x)}\, dx \\
&= e^{-n\, c(\eta)} \int_{\mathcal{X}} h(x) e^{(\eta+s)\cdot T(x)}\, dx \\
&= e^{n\,[c(\eta+s) - c(\eta)]},
\end{aligned}$$

so we can find its mean and (co)variance as

$$\begin{aligned}
\mathsf{E}[T] &= \nabla_\eta \log M_T(0) &= n\, \nabla_\eta c(\eta) \\
\mathsf{V}[T] &= \nabla^2_\eta \log M_T(0) &= n\, \nabla^2_\eta c(\eta).
\end{aligned}$$

3.1. The Information Inequality

The Fisher Information about the natural parameter η from a single observation ($n = 1$) from an exponential family $f(x \mid \eta) = h(x) \exp\big(\eta \cdot T(x) - c(\eta)\big)$ is given by

$$I(\eta) = -\nabla_\eta^2 \log f(x \mid \eta) = \nabla_\eta^2 c(\eta),$$

so by Equation (1) the information in any parametrization is given by

$$I(\theta) = J^\intercal \, \nabla_\eta^2 c(\eta) \, J$$

for $J = \partial\eta/\partial\theta$. The natural sufficient statistic $T(x)$ has mean $\psi(\theta) \equiv \mathsf{E}[T(X) \mid \theta] = \nabla_\eta c(\eta(\theta))$.

In particular, for scalar ($p = 1$) exponential families, the Information Inequality takes the form

$$\begin{aligned}
\mathsf{V}[T] &\geq \frac{|\psi'(\theta)|^2}{I(\theta)} \\
&= \frac{|c''(\eta(\theta))\,\eta'(\theta)|^2}{\eta'(\theta)c''(\eta)\eta'(\theta)} \\
&= c''(\eta(\theta)) \\
&= \mathsf{V}[T],
\end{aligned}$$

so the lower bound is attained. This is not terribly surprising, since the inequality was based on the covariance inequality for the random variables T and $Z \equiv \nabla_\theta \log f(x \mid \theta) = \nabla_\theta \eta(\theta) \cdot T(X) - \nabla_\theta c(\theta)$, which are related by an affine transformation for scalar exponential families and hence are perfectly correlated.

4. Objective Bayesian Analysis

Laplace in the 1700's used the *uniform* prior distribution $\pi(\theta) \equiv 1$ in his Bayesian statistical analysis, intending it to represent a complete absence of knowledge about θ before observing a data vector $x \in \mathcal{X}$, and leading to a posterior density function

$$\pi(\theta \mid x) \propto f(x \mid \theta)$$

proportional to the likelihood. As appealing as this is for a non-subjective analysis, it is not invariant under reparametrization; for example, if we use the uniform distribution $\pi_1(\theta) \equiv 1$ for a binomial success probability $\theta \in \Theta = (0, 1)$ then upon observing y successes in n tries this leads to

$$\pi_1(\theta \mid x) \propto \theta^y (1-\theta)^{n-y}$$

the $\mathsf{Be}(1 + y, 1 + n - y)$ posterior distribution, with mean $T_1(Y) = \mathsf{E}^{\pi_1}[\theta \mid y] = \frac{1+y}{2+n}$, while a similar analysis using a uniform prior density for the (natural) logistic parameter $\eta = \log \frac{\theta}{1-\theta}$ leads to the beta $\mathsf{Be}(y, n - y)$ posterior distribution, with mean $T_2(Y) = \mathsf{E}^{\pi_2}[\theta \mid y] = \frac{y}{n}$. When η is accorded a uniform prior the *implicit* prior for θ is $\pi_2(\theta) = 1(\eta)\eta'(\theta) = \frac{1}{\theta(1-\theta)}$, the beta $\mathsf{Be}(0, 0)$, while the uniform density $\pi_1(\theta) \equiv 1$ is also the beta $\mathsf{Be}(1, 1)$. For large numbers of success y and failure $n - y$ these two *reference* posterior distributions are very close, but for small y or $n - y$ they are not; which should we use, and why?

Evidently the problem is that in transforming from η to θ any prior density $\pi^\eta(\eta)$ will be transformed to $\pi^\theta(\theta) = |\det J|\pi^\eta(\eta(\theta))$ where $J = \partial\eta/\partial\theta$; the functional form of the priors in these two parametrizations cannot be the same, because of the Jacobian. Harold Jeffreys noticed that, since the Fisher Information $I(\theta) = J^\intercal\, \nabla^2_\eta c(\eta)\, J$ transforms bilinearly in J, the recipe

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)}$$

will give a prior density that transforms consistently to any parametrization:

$$\begin{aligned}
\pi_J(\theta) &\propto \sqrt{\det I(\theta)} \\
&= \sqrt{\det J^\intercal\, I(\eta)\, J} \\
&= |\det J|\, \sqrt{\det I(\eta)} \\
&= |\det J|\, \pi_J(\eta),
\end{aligned}$$

as required. In the case of Binomial data, for example, $I(\theta) = \frac{1}{\theta(1-\theta)}$ so $\pi_J(\theta) \propto \theta^{\frac{1}{2}-1}(1-\theta)^{\frac{1}{2}-1}$ is the $\mathsf{Be}(\frac{1}{2}, \frac{1}{2})$ distribution (also called the *arc-sin* distribution, since it has CDF $\mathsf{P}[\theta < t] = \frac{2}{\pi}\sin^{-1}(\sqrt{t})$), leading to a $\mathsf{Be}(\frac{1}{2} + y, \frac{1}{2} + n - y)$ posterior distribution for the success probability θ upon observing y successes in n tries, with posterior mean $\mathsf{E}[\theta \mid y] = (y+\frac{1}{2})/(n+1)$.