17
Lindley Information Measure

In the last Lecture we have presented and discussed Degroot's results concerning measuring information as a change in utility. We have that if the decision maker's utility is the log of a certain function of the parameters in a model, then the Bayesian choice would be the most recent distribution of the parameters, usually the posterior distribution.

In this Lecture we will get into the details of a particular case of Degroot's result, the so called Lindley information measure (Lindley (1956)[4]). His basic theorems will be presented here in the format of propositions instead, without any proof, though.

We will explore two examples presented in Parmigiani and Berry (1994) [8] and related to design of clinical experiments.

17.1 Terminology and basic results

From the point of view of decision making, information is anything that enables us to make a better decision, that is a decision with a higher expected utility. The expected change in utility can actually be used as a quantitative measure of the worth of an experiment in any given situation. It is natural to think that a some point the experimentation must be stopped because a certain amount of information has been attained.

Lindley's measure of information is a well known measure that pursue such natural concepts. Below, we define it and present some of its basic properties, followed by two examples.

Consider the decision problem of reporting a distribution regarding an unknown quantity θ with values in Θ. This is a Bayesian way of modeling a situation in which data are not gathered to solve a specific decision problem but rather to learn about the world or to provide several decision makers with useful information in different decision problems.

Let the reported distribution be φ, and let the utility function be $U(\theta, \varphi) = \log \varphi(\theta)$. The Bayes decision is $\varphi = \pi$, where π is the current distribution on θ, as has been shown to be a proper local scoring rule. The value associated with the decision problem is:

$$\mathcal{I}_0 \equiv E\{\log \pi(\Theta)\} = \int_\Theta \log(\pi(\theta))\pi(\theta)d\theta \tag{17.1.1}$$

Assume now that the outcome of the experiment $\mathcal{E}$, consisting of observing X from the distribution $p(x|\theta)$, becomes known. By the same argument as above, the value associated with the decision problem is:

$$\mathcal{I}_1(x) \equiv E\{\log \pi(\Theta|X = x)|X = x\}.$$

The expected worth of the experiment $\mathcal{E}$ before performing it, is

$$\begin{aligned} \mathcal{I}(\mathcal{E}) &\equiv E\{\log \pi(\Theta|X)\} - E\{\log \pi(\Theta)\} \qquad (17.1.2) \\ &= E\left\{\log\left(\frac{\pi(\Theta|X)}{\pi(\Theta)}\right)\right\} \\ &= \int_\Theta \int_\mathcal{X} \log\left[\frac{\pi(\theta|x)}{\pi(\theta)}\right] p(x|\theta)\pi(\theta)dxd\theta \end{aligned}$$

Lindley's measure of information was directly related to information theoretic arguments (see also Kullback and Leibler (1951) [3] and Kullback (1959) [2]. We will now present, without proof, some of the basic results related to the Lindley's measure of information.

Proposition 17.1.1 $\mathcal{I}(\mathcal{E}) \geq 0$, *with equality if, and only if,* $p(x|\theta)$ *does not depend on* Θ.

Note that the observed information is not necessarily nonnegative, and the experiment might result in lost of information due to, for instance, unexpected values of x.

Suppose, for now, that the experiment $\mathcal{E}$ consists of observing x_1 and x_2, and that the partial experiments can be represented by $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively. The next results says that the overall information in both experiments can be represented by the information contained in the first one plus the additional information gained by the second one.

Proposition 17.1.2 $\mathcal{I}(\mathcal{E}) = \mathcal{I}(\mathcal{E}_1) + \mathcal{I}(\mathcal{E}_2|\mathcal{E}_1)$.

If the second experiment does not have any additional information, that is $\mathcal{I}(\mathcal{E}_2|\mathcal{E}_1) = 0$, we have the following result.

Proposition 17.1.3 *If x_1 is sufficient for x in the Neyman-Fisher sense, then $\mathcal{I}(\mathcal{E}) = \mathcal{I}(\mathcal{E}_1)$.*

Suppose that x_1 and x_2 are independents, conditional on θ, that is $p(x_1, x_2|\theta) = p(x_1|\theta)p(x_2|\theta)$. In such situations it is natural to say that $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, and the following result holds.

Proposition 17.1.4 *If $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent*

$$\mathcal{I}(\mathcal{E}_2|\mathcal{E}_1) \leq \mathcal{I}(\mathcal{E}_2) \tag{17.1.3}$$

with equality if, and only if, x_1 and x_2 are unconditionally independent.

Basically, what the proposition says is that the information of a particular experiment would be decreased (on average) if it is performed after any other independent experiment. Consequently, we have the following result based on the previous one,

Proposition 17.1.5 *If $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent experiments, then*

$$\mathcal{I}(\mathcal{E}_1) + \mathcal{I}(\mathcal{E}_2) \geq \mathcal{I}(\mathcal{E}) \tag{17.1.4}$$

with equality if, and only if, x_1 and x_2 are unconditionally independent.

In the case of identical experiments being repeated sequentially a more general result on the lines of the last proposition can be stated. Let $\mathcal{E}^{(1)} = \mathcal{E}_1$ be any experiment and let $\mathcal{E}_2, \mathcal{E}_3, \cdots$ be independent and identical repetitions of $\mathcal{E}^{(1)}$. Let also, $\mathcal{E}^{(n)} = (\mathcal{E}_n, \mathcal{E}^{(n-1)})$.

Proposition 17.1.6 *$\mathcal{I}(\mathcal{E}^{(n)})$ is a concave, increasing function of n.*

Another property of $\mathcal{I}(\mathcal{E})$ is stated in the follwing proposition

Proposition 17.1.7 *For fixed $\mathcal{E}$, $\mathcal{I}(\mathcal{E})$ is a concave function of $\pi(\theta)$.*

We will now discuss a particular situation where the experiment $\mathcal{E}$ can be represented by a design variable D, where D can be, for instance, the sample size for a specific experiment, or even the situation where a continuous variable cannot be observed exactly, but we know whether or not such variable is greater than some point D. In such cases, if we add some mild regularity conditions the following result might be useful to help finding the optimal design, that is the design that maximizes the expected utility.

17.2 First-Order Conditions

Consider a family of experiments $\mathcal{E}_D$ indexed by the design variable D. The likelihood function for the experiment $\mathcal{E}_D$ is $p_D(x|\theta)$, $x \in \mathcal{X}$. Let $\pi_D(\theta|x)$ be the posterior distribution resulting from the observation of $X = x$ and let

$m_D(x)$ the marginal distribution of X. The expected information is then given by:

$$\mathcal{I}(D) = \int_\Theta \int_\mathcal{X} \log\left[\frac{\pi_D(\theta|x)}{\pi(\theta)}\right] p_D(x|\theta)\pi(\theta)dxd\theta \tag{17.2.1}$$

Let us assume that D is one-dimensional and continuous or approximable by a continuous variable. We also assume that $p_D(\theta|x)$ is differentiable with respect to D, so

$$p'_D(x|\theta) \equiv \frac{\partial p_D(x|\theta)}{\partial D} \qquad\qquad \pi'_D(\theta|x) \equiv \frac{\partial \pi_D(\theta|x)}{\partial D}$$

Theorem 17.2.1 *The derivative of $\mathcal{I}(D)$ with respect to D is*

$$\frac{d\mathcal{I}(D)}{dD} = \int_\Theta \int_\mathcal{X} \log\left[\frac{\pi_D(\theta|x)}{\pi(\theta)}\right] p'_D(x|\theta)\pi(\theta)dxd\theta. \tag{17.2.2}$$

Proof: To see that (17.2.2) holds, differentiate the integrand of (17.2.1) to get:

$$\left[\log\left[\frac{\pi_D(\theta|x)}{\pi(\theta)}\right] p'_D(x|\theta) + p_D(x|\theta)\frac{\pi'_D(\theta|x)}{\pi_D(\theta|x)}\right]\pi(\theta).$$

From Bayes' rule

$$\pi'_D = \pi\frac{p'_D m_D - p_D m'_D}{m_D^2}.$$

Thus

$$\frac{p_D\pi'_D}{\pi_D} = \frac{m_D\pi'_D}{\pi} = \frac{p'_D m_D - p_D m'_D}{m_D} = p'_D - p_D\frac{m'_D}{m_D}.$$

Now from the area condition,

$$\begin{aligned} \int_\Theta \int_\mathcal{X} p_D(x|\theta)\pi(\theta)\frac{m'_D(x)}{m_D(x)}dxd\theta = \int_\mathcal{X} m'_D(x)dx &= 0, \\ \int_\Theta \int_\mathcal{X} p'_D(x|\theta)\pi(\theta)dxd\theta &= 0, \end{aligned}$$

And (17.2.2) follows.

With simple manipulations, (17.2.2) can be written:

$$\frac{d\mathcal{I}(D)}{dD} = \int_\Theta \int_\mathcal{X} \log\left[p_D(x|\theta)\right] p'_D(x|\theta)dxd\theta - \int_\mathcal{X} \log\left[m_D(x)\right] m'_D(x)dx. \tag{17.2.3}$$

This expression parallels the alternative expression for (17.2.1), given by Lindley (1956):

$$\mathcal{I}(D) = \int_\Theta \int_\mathcal{X} \log\left[p_D(x|\theta)\right] p_D(x|\theta)dxd\theta - \int_\mathcal{X} \log\left[m_D(x)\right] m_D(x)dx. \tag{17.2.4}$$

Formulas (17.2.2) and (17.2.3) can be used in calculating the optimal D whenever Lindley information enters as a part of the design criterion.

If D is sample size, and the cost of observation c is fixed, then the real valued approximation to the optimal sample size satisfies

$$\frac{d\mathcal{I}(D)}{dD} - c = 0. \tag{17.2.5}$$

In realistic applications closed form expressions form for $\mathcal{I}$, (17.2.2) or (17.2.3) are hard to derive. Müller and Parmigiani (1993,1995) [5], [6] discuss the use of Markov-chain Monte Carlo methods to evaluate information theoretic measures such as entropy and Lindley information and introduce fast algorithms for optimization problems that arise in Bayesian design. Bielza, Müller and Ríos Insua (1998) [1] is another interesting work related to MCMC methods in Decision Analysis.

17.3 Examples

17.3.1 Example 1: Partitioning Y

An interesting special case arises when $\mathcal{I}(D)$ has a maximum. Consider for example finding the most informative binary partition for an observation Y. This problem arises in applications when a continuous variable has to be turned into a categorical (in this case binary) variable. So let the random variable Y have conditional density $f(y|\theta)$. We are interested in learning about Θ. Assume that y cannot be observed exactly: we may only observe whether or not it is greater than some point D. What is the best D? Define X to be a Bernoulli random variable with success probability $F(D|\theta)$. Using (17.2.3), a necessary condition for D to be optimal is:

$$\int_\Theta \log\left[\frac{F(D|\theta)}{1-F(D|\theta)}\right]\pi(\theta|y=D)d\theta = \log\frac{F(D)}{1-F(D)}; \tag{17.3.1}$$

in words, we must choose the design so that the expected conditional logit is equal to the marginal logit.

As an example, take Y to be exponential with unknown failure rate θ. Let the prior of $tTheta$ be a conjugate Gamma(α, β). Then, using (17.3.1) the optimal D is the unique solution of:

$$\beta^{\alpha+1}\zeta(\alpha+2)\frac{e^{1+\beta/D}}{D^{\alpha+1}} = \log\left[\left(\frac{\beta+D}{\beta}\right)^{\alpha} - 1\right]$$

where ζ is Riemann's zeta function. Because the left hand side is always positive, the optimal D is greater that the marginal median of Y.

An alternative approach to the derivation of the same equations, based on a Kullback–Leibler divergence between the actual posterior distribution and the posterior distribution that would be available if Y could be observed exactly, is discussed in Parmigiani (1992) [7].

17.3.2 Example 2: Information and Sample Size : The Exponential case

A common design problem in clinical and other applications is determining the best fixed sample size for the experiment. To make the notation more standard, let $D = n$. If the data to be collected are exchangeable, proposition 17.1.6 applies to show that $\mathcal{I}(n)$ is increasing and convex in n. Therefore, if the cost of observation increases linearly or faster in n and if information and cost are additive, there are at most two (adjacent) solutions.

We will consider here the exponential survival data with unknown failure rate Λ and conjugate prior gamma(α, β). Let $t = \sum_i^n x_i$. The expected information in a sample of size n is:

$$\begin{aligned} \mathcal{I}(n) &= \int\int \log\left[\frac{\Gamma(\alpha)(\beta+t)^{\alpha+n}}{\Gamma(\alpha+n)\beta^\alpha}\lambda^n e^{-\lambda t}\right]\frac{\beta^\alpha}{\Gamma(\alpha)\Gamma(n)}\lambda^{\alpha+n-1}t^{n+1}e^{-\lambda(b+t)}dtd\lambda \\ &= \log\frac{\Gamma(a)}{\Gamma(a+n)} + (a+n)\psi(a+n) - a\psi(a) - n \qquad (17.3.2) \end{aligned}$$

where ψ is the digamma function and the integral are from zero to infinity.

The value of $\mathcal{I}(n)$ does not depend on β. The amount one expects to learn depends solely on the number of prior failures and not on the prior total survival.

A useful corollary of (17.3.2) is the expected information from the next observation, which is

$$\mathcal{I}(1) = \frac{1}{\alpha} + \psi(\alpha) - \log\alpha.$$

If $\alpha = 1$, then $\mathcal{I}(1) = 1 - C = .4288$, where C is Euler's constant.

The following large-n approximation brings out further interesting features of $\mathcal{I}(n)$. Using

$$\log\Gamma(n) \approx \frac{1}{2}\log 2\pi - n + (n - \frac{1}{2})\log n$$

and

$$\psi(n) \approx \log n - \frac{1}{2n},$$

we get:

$$\mathcal{I}(n) \approx K(\alpha) + \frac{1}{2}\log(\alpha + n)$$

where $K(\alpha) = \log\Gamma(\alpha) - \alpha(\psi(\alpha) - 1) - \frac{1}{2}\log 2\pi - \frac{1}{2}$. Using this approximation one can solve analytically for the optimal fixed sample size with sampling cost c (in information units) per observation. In particular, taking n to be continuous, the optimum is:

$$n^* = \frac{1}{2c} - a.$$

References

[1] C. Bielza, P. Mueller, and D. Insua. Monte carlo methods for decision analysis with applications to influence diagrams. 1998.

[2] S. Kullback. *Information Theory and Statistics*. New York, 1959.

[3] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[4] D. V. Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

[5] P. Müller and G. Parmigiani. Numerical evaluation of information theoretic measures. In *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*, 1993.

[6] P. Müller and G. Parmigiani. Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association*, 90:1322–1330, 1995.

[7] G. Parmigiani. Design of partially observable survival data. Discussion Paper 92-A34, Institute of Statistics and Decision Sciences, 1992.

[8] G. Parmigiani and D. A. Berry. Applications of Lindley information measure to the design of clinical experiments. In *Aspects of Uncertainty. A Tribute to D. V. Lindley*, pages 351–362. John Wiley & Sons, 1994.