Inversion of Bayes Formula and measures of Bayesian information gain and pairwise dependence

Kai Wang Ng
The University of Hong Kong

and Howell Tong
London School of Economics and Political Science
& The University of Hong Kong

Abstract. By inverting the Bayes formula in a point-wise manner, we develop measures quantifying the information gained by the Bayesian process, in reference to the Fisher information. Simple examples are used for focused illustrations of the ideas. Numerical computation for the measures is discussed with formulae. By extending the information gain concept to the broader context of distribution theory, we arrive at a pairwise dependence measure, which can handle the case of functional dependence and becomes Pearson's ϕ^2 when the joint probability density function (pdf) is defined.

Keywords: Likelihood; Bayes Formula; Inverse Bayes Formula; Bayesian information gain function; information gain index; pairwise dependence measure ψ^2; pairwise dependence index; Pearson's ϕ^2.

1. Introduction

In standard Bayesian notation, we use $\pi(\theta)$ to denote the prior probability density function (pdf) of parameter θ with support $\mathcal{S}(\Theta)$, $L(y|\theta)$ the likelihood function (i.e. the pdf of data given the parameter) with support $\mathcal{S}(Y|\theta)$, $p(\theta|y)$ the posterior pdf with support $\mathcal{S}(\Theta|y)$ of parameter given the data, and $f(y)$ the unconditional pdf for the data with support $\mathcal{S}(Y)$. Both θ and y can be vectors. Note that in general, the projection of $\mathcal{S}(Y|\theta)$ into $\mathcal{S}(Y)$ is a subset, i.e. $\mathcal{S}(Y|\theta) \overset{pj.}{\subset} \mathcal{S}(Y)$, and the equality $\mathcal{S}(Y|\theta) \overset{pj.}{=} \mathcal{S}(Y)$ may hold for some θ. In regard to integral or probability, the latter is essentially the same as when the complement of the projection of $\mathcal{S}(Y|\theta)$ into $\mathcal{S}(Y)$ is a set of measure zero. If the joint support $\mathcal{S}(\Theta, Y)$ equals the product space $\mathcal{S}(\Theta) \times \mathcal{S}(Y)$, then $\mathcal{S}(Y|\theta) \overset{pj.}{=} \mathcal{S}(Y)$ for all θ; and vice versa. A similar relationship is true between $\mathcal{S}(\Theta|y)$ and $\mathcal{S}(\Theta)$.

From the joint pdf identity, $L(y|\theta)\pi(\theta) = p(\theta|y)f(y)$, the Bayes formula

$$p(\theta|y) = \pi(\theta)L(y|\theta)/\int_{\mathcal{S}(\Theta|y)} \pi(\theta)L(y|\theta)d\theta,$$

follows by a substitution of $f(y)$, which is expressed as the integral of the joint pdf with respect to θ over $\mathcal{S}(\Theta|y)$. We can re-write the above joint pdf identity as $\pi(\theta)L(y|\theta)/p(\theta|y) = f(y)$, where (θ, y) is in the joint support $\mathcal{S}(\Theta, Y)$. Now for any fixed θ, we can integrate both sides of the re-expressed joint pdf identity with respect to y over $\mathcal{S}(Y|\theta)$ and obtain the prior pdf at θ,

$$\begin{aligned} \pi(\theta) &= \int_{\mathcal{S}(Y|\theta)} f(y)dy \left\{ \int_{\mathcal{S}(Y|\theta)} \frac{L(y|\theta)}{p(\theta|y)} dy \right\}^{-1} && (1.1) \\ &\leq \left\{ \int_{\mathcal{S}(Y|\theta)} \frac{L(y|\theta)}{p(\theta|y)} dy \right\}^{-1}, && (1.2) \end{aligned}$$

where the equality holds if and only if $\mathcal{S}(Y|\theta) \stackrel{pj.}{=} \mathcal{S}(Y)$, or the complement of the projection of $\mathcal{S}(Y|\theta)$ into $\mathcal{S}(Y)$ is a set of measure zero. In particular, under the so-called "positivity assumption", (cf. Tanner and Wong, 1987; and Tanner, 1996, Chapter 5), where $\mathcal{S}(\Theta, Y) = \mathcal{S}(\Theta) \times \mathcal{S}(Y)$, we have

$$\pi(\theta) = \left\{ \int_{\mathcal{S}(Y|\theta)} \frac{L(y|\theta)}{p(\theta|y)} dy \right\}^{-1}, \quad \forall \theta \in \mathcal{S}(\Theta). \tag{1.3}$$

In the words of Meng (1996, p.311), the explicit form (1.3) 'was "mysteriously" missing in the general literature.' This may be due to the tradition in the Bayesian literature to express the posterior distribution in terms of the prior distribution. We shall follow Ng (1995, 1997) and call (1.3) the (point-wise) Inverse Bayes Formula (IBF), in order to emphasize its unconventional character, in that the prior distribution is expressed in terms of the the posterior distribution. In fact, it is the harmonic mean of $p(\theta|y)$ with respect to $L(y|\theta)$.

The not-so-well-known (1.3) deserves to be better known because it can lead to a number of important consequences as those already discussed in the above-cited papers and in Tan, Tian and Ng (2009). The objective of this paper is to continue the exploration of other consequences, including some unexpected.

The plan of our paper is as follows. In Section 2, we introduce two natural functions measuring Bayesian information gain, in reference to Fisher's information function, and justify them with the aid of (1.2). And we propose two normalized information gain indices between 0 and 1, Γ_π and $\Gamma_\pi(y)$, the former measuring the total information gain aggregated over all parameter values and all possible data and the latter measuring the information gain on all parameter values for the datum y at hand. After fixing the ideas with simple examples, we discuss numerical computation. In Section 3, we extend the concept of total Bayesian information gain to the context of distribution theory, yielding a natural pairwise dependence measure, ψ^2, between two random variables or random vectors under conditional specification. This measure is normalized to a pairwise dependence index, δ, which takes a value between zero and unity. The index δ equals zero if and only if we have an independent pair. It equals unity for a highly dependent pair, which can even be functionally dependent. When the unconditional joint pdf is defined, ψ^2 is the same as Pearson ϕ^2. However, unlike Pearson's ϕ^2, our ψ^2 does not require the existence of the joint pdf. The difference is demonstrated by a functional dependence example, for which ϕ^2 is not defined. To illustrate the main ideas, we include what we hope are instructive examples. In Section 4, we draw some conclusions and describe an alternative sensitivity function.

2. Information Gain

The Fisher information function, $I(\theta)$, measures one particular kind of information regarding the parameter in the sense that if $I(\theta_1)$ is larger than $I(\theta_2)$, then the precision of likelihood inference is higher at θ_1 than at θ_2. It is defined as the expectation of the squared derivative of the log-likelihood function, where the expectation is taken with respect to the pdf of the data *with the same θ as used in the likelihood*. There are two essential ingredients in Fisher's construction: (i) the function (of the parameter and the data) to be aggregated and (ii)

the distribution with which to perform the aggregation. The function in (i) should reflect the sensitivity to changes in the parameter. As a measure of inference precision whose inverse measures the variability (or uncertainty of inference), the positive definiteness of the resulting function upon aggregation is particularly required. For want of better terms, we shall call the function in (i) the sensitivity function and the distribution in (ii) the aggregating distribution.

Similarly, in quantifying the information gained by a Bayesian process, the pertinent question is *what sensitivity function* should be aggregated with respect to *which aggregating distribution*. Since the input to a Bayesian inferential process is the prior pdf, $\pi(\theta)$, and the output from it is the posterior pdf, $p(\theta|y)$, in the light of an observation y, it is natural to consider the change $(p(\theta|y) - \pi(\theta))$ as the sensitivity function. Following Fisher, it is natural to use the likelihood $L(y|\theta)$ as the aggregating distribution (over all possible data). The choice entails the following *Bayesian Information Gain on Parameter* relative to the prior π:

$$BIGP_{\pi}(\theta) \equiv E_{L(y|\theta)}[p(\theta|y) - \pi(\theta)] = E_{L(y|\theta)}[p(\theta|y)] - \pi(\theta). \quad (2.1)$$

Here and later, we use the notation E_{ν} to denote the expectation with respect to the pdf ν. Note that the first term on the right-hand side of (2.1) is not difficult to obtain by simulation if the variate y given θ can be generated. The subscript π in $BIGP_{\pi}(\theta)$, which indicates the dependence of the information gain function on the choice of the prior distribution π, can be dropped whenever the context is clear. Clearly, before we can accept $BIGP_{\pi}(\theta)$ as a measure it must be non-negative, since $\pi(\theta)$ reflects the available information about θ without any data. We now show by means of (1.2) that $BIGP_{\pi}(\theta) \geq 0$ always.

Let AM and HM be respectively the arithmetic mean and the harmonic mean of a function of a random variable with respect to the

distribution of the random variable. It is well known that

$$AM \geq HM. \tag{2.2}$$

Treating the posterior pdf as a function of the data and the likelihood as the pdf of the data and by virtue of (1.2), we can express (2.1) as

$$\begin{aligned} BIGP_\pi(\theta) &= \int_{\mathcal{S}(Y|\theta)} p(\theta|y)L(y|\theta)dy - \pi(\theta) \qquad (2.3)\\ &\geq \int_{\mathcal{S}(Y|\theta)} p(\theta|y)L(y|\theta)dy - \left\{\int_{\mathcal{S}(Y|\theta)} \frac{L(y|\theta)}{p(\theta|y)}dy\right\}^{-1} \\ &= AM - HM \geq 0, \end{aligned}$$

where the equality in the second step holds if and only if $\mathcal{S}(Y|\theta) \overset{pj.}{=} \mathcal{S}(Y)$, or the complement of the projection of $\mathcal{S}(Y|\theta)$ into $\mathcal{S}(Y)$ is a set of measure zero. In the last step, the equality $BIGP_\pi(\theta) = 0$ holds for all θ if and only if y and θ are independent, i.e. no information is gained if y carries no information about the parameter. Since $BIGP_\pi(\theta)$ measures the information gain at each θ and is always non-negative, we may consider the *total Bayesian Information Gain*, or BIG_π for short, by integrating the function $BIGP_\pi(\theta)$ (or summing if we are dealing with a discrete distribution):

$$BIG_\pi \equiv \int_{\mathcal{S}(\Theta)} BIGP_\pi(\theta)d\theta = \int_{\mathcal{S}(\Theta)} \left\{\int_{\mathcal{S}(Y|\theta)} L(y|\theta)p(\theta|y)dy\right\} d\theta - 1. \tag{2.4}$$

Note that BIG_π is invariant with respect to a one-to-one transformation of the data y as well as the parameter θ. Now, in the repeated integral, the outside one does not have the interpretation of an expectation as the inside one does, and thus the result may be positive infinity. Thus it is often more convenient to use the following normalized form, called the *Information Gain Index*, or Γ_π for short, which

is confined to the closed unit interval $[0, 1]$:

$$\Gamma_\pi \equiv BIG_\pi/(1 + BIG_\pi) = 1 - \left\{ \int_{\mathcal{S}(\Theta)} \left\{ \int_{\mathcal{S}(Y|\theta)} L(y|\theta)p(\theta|y)dy \right\} d\theta \right\}^{-1}. \tag{2.5}$$

The index is 0 if the parameter and the data are independent, and 1 if the total information gain is positive infinity. Different choices of the prior distribution $\pi(\theta)$ can be compared by reference to Γ_π, as demonstrated later in examples.

A Bayesian data analyst may sometimes find the information gain for data at hand more attractive than the total information gain aggregated over all possible data. In that case, it is more relevant to use the following *Bayesian Information Gain conditional on the data*,

$$BIGD_\pi(y) \equiv E_{p(\theta|y)}L(y|\theta) - E_{\pi(\theta)}L(y|\theta) = E_{p(\theta|y)}L(y|\theta) - f(y) \tag{2.6}$$

and the corresponding *Bayesian Information Gain Index conditional on the data*,

$$\Gamma_\pi(y) \equiv BIGD_\pi(y)/(1 + BIGD_\pi(y)), \tag{2.7}$$

which obviously involves less and easier calculation, especially for y in a numerical form. Using a completely analogous argument for $BIGP_\pi(\theta)$ in (2.3), we can show that $BIGD_\pi(y) \geq 0$ always, confirming its legitimacy as an information gain measure.

We can also get BIG_π through $BIGD_\pi(y)$,

$$BIG_\pi = \int_{\mathcal{S}(Y)} BIGD_\pi(y)dy = \int_{\mathcal{S}(Y)} \left\{ \int_{\mathcal{S}(\Theta|y)} L(y|\theta)p(\theta|y)d\theta \right\} dy - 1, \tag{2.8}$$

provided that the order of integration can be interchanged. However, there is no direct relationship between $\Gamma_\pi(y)$ defined in (2.7) and the Γ_π defined in (2.5); specifically

$$\Gamma_\pi \neq \int_{\mathcal{S}(Y)} \Gamma_\pi(y)dy. \tag{2.9}$$

We demonstrate the ideas by three simple examples before considering practical computation at the end of this section.

Example 1: For the Binomial likelihood

$$L(y|\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y},$$

the usual prior distribution is from the Beta family, Be(a,b), of conjugate priors with pdf

$$\pi(\theta) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b),$$

where $B(a,b)$ is the Beta function. The posterior pdf is then

$$p(\theta|y) = \theta^{y+a-1}(1-\theta)^{n-y+b-1}/B(y+a, n-y+b)$$

The Bayesian information gain on parameter is

$$BIGP_\pi(\theta) = \sum_{y=0}^{n} \frac{\left(\begin{array}{c} n \\ y \end{array} \right) \theta^{2y+a-1}(1-\theta)^{2n-2y+b-1}}{B(y+a, n-y+b)} - \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

and the total Bayesian information gain is thus

$$BIG_\pi = \sum_{y=0}^{n} \left(\begin{array}{c} n \\ y \end{array} \right) \frac{B(2y+a, 2(n-y)+b)}{B(y+a, n-y+b)} - 1,$$

Since the uniform prior with $a = b = 1$ is commonly used whenever there is no information on θ before collecting the data, intuition would suggest that this prior should result in the maximum information gain after the Bayesian processing. It can be shown, however, that BIG_π as well as Γ_π is a decreasing function of a, b. That is, the prior with $a < 1$ and $b < 1$ yields larger information gain than the uniform prior. This includes Jeffreys prior with $a = b = 0.5$. Furthermore, when a or b approaches 0, BIG_π approaches ∞ and hence the normalized index Γ_π approaches 1.

On the other hand, BIG_π is an increasing function of n, confirming the intuition that a larger sample size n should yield a larger information gain.

Instead of the aggregated information gain, we may sometimes wish to focus on the information gain at a particular observation y. In this case, we first note the unconditional pdf for the observation y:

$$g(y) = \left(\begin{array}{c} n \\ y \end{array} \right) \frac{B(y+a, n-y+b)}{B(a,b)}$$

and obtain the Bayesian information gain conditional on the observation y at hand, namely

$$BIGD_\pi(y) = \left(\begin{array}{c} n \\ y \end{array} \right) \left(\frac{B(2y+a, 2(n-y)+b)}{B(y+a, n-y+b)} - \frac{B(y+a, n-y+b)}{B(a,b)} \right).$$

As a function of a and b, $BIGD_\pi(y)$ depends on both y and n in a much more complicated way, but it can be computed easily using the software SAS or R.

Example 2: Consider the following likelihood function as a dislocated exponential with a positive but unknown location parameter and an exponential prior pdf for the positive parameter,

$$L(y|\theta) = e^{-(y-\theta)} \text{ for } y > \theta > 0; \quad \pi(\theta) = \lambda\, e^{-\lambda\theta} \text{ for } \theta > 0, \quad \text{where } \lambda > 0.$$

Since λ completely determines the prior distribution in this family, we wish to find the λ which achieves the maximum of information gain Γ_λ. In this case, the posterior pdf is defined in the domain $0 < \theta < y$:

$$p(\theta|y) = 1/y \text{ if } \lambda = 1, \quad (\lambda-1)e^{-(\lambda-1)\theta}/(1-e^{-(\lambda-1)y}) \text{ if } \lambda \neq 1.$$

In words, if $\lambda = 1$, the posterior pdf is uniform in $(0, y)$. If $\lambda > 1$, it is a right-truncated Exponential$(\lambda - 1)$ defined in $(0, y)$. If $\lambda < 1$, it is proportional to the increasing function of θ, $(1-\lambda)e^{(1-\lambda)\theta}$, but normalized within the interval $(0, y)$.

For $\lambda \leq 1$, the repeated integral in (2.5) is positive infinity and thus the information gain index $\Gamma_\lambda = 1$. We can easily use any software to compute Γ_λ for $\lambda > 1$, obtaining the following table and Figure 1:

λ	1	2	3	4	5	6	7	8	9	10
Γ_λ	1	0.392	0.279	0.217	0.178	0.151	0.131	0.116	0.104	0.094

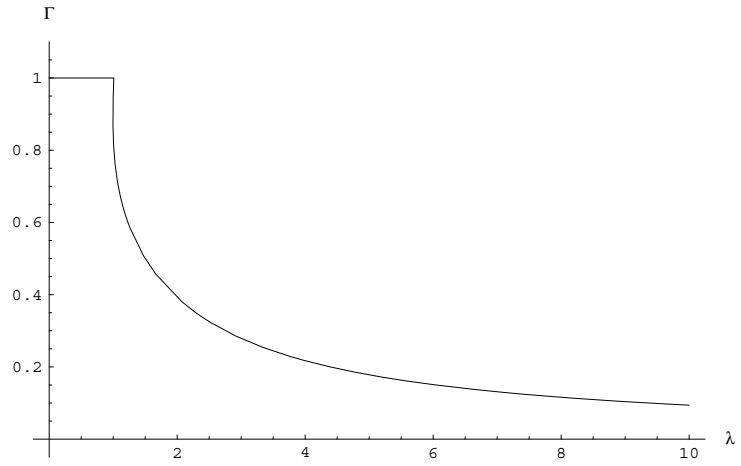

Figure 1: Information Gain Index as a function of λ

In conclusion, the measure Γ_λ achieves the maximum possible value of 1, if $\lambda \leq 1$. This is quite natural since λ equals the modal value of pdf at the origin. The greater its value, the greater the concentration around the origin, so that a large λ corresponds to high prior information to start with, leading to a correspondingly small gain in the end.

For completeness of the example, the unconditional pdf of the data, $y > 0$, is as follows:

$$g(y) = ye^{-y} \text{ if } \lambda = 1, \quad (\lambda/(\lambda-1))e^{-y}(1 - e^{-(\lambda-1)y}) \text{ if } \lambda \neq 1.$$

Thus, if $\lambda = 1$, $g(y)$ is Gamma(2); otherwise it is a mixture of Exponential(1) and Exponential(λ) with respective weights of w and $1-w$, where $w = \lambda/(\lambda-1)$ for $\lambda > 1$ and $w = 1 - (1-\lambda)^{-1}$ for $\lambda < 1$.

Example 3: Consider the following likelihood, which is not so standard,

$$L(y|\theta) = e^{|\theta|}(2\pi y)^{-1/2}\exp(-\theta^2/2y - y/2), \quad y > 0,\ -\infty < \theta < \infty.$$

This likelihood is related to the inference of the mean velocity θ (negative sign for moving towards left) of a particle moving in a linear Brownian motion. Other parameters have been omitted for a simpler illustration here. The sample average of first-passage times of such a particle over a unit length of distance is distributed as an inverse Gaussian distribution and the reciprocal of the sample average is interpreted as an estimator of the velocity of the particle. See Johnson *et al.* (1994, Chapter 15) for more detail.

As suggested by the range of θ, suppose that we wish the posterior distribution of θ to be as simple as the normal with zero mean and variance y, $N(0, y)$, i.e.

$$p(\theta|y) = (2\pi y)^{-1/2}\exp(-\theta^2/2y), \quad -\infty < \theta < \infty,\ y > 0.$$

However, in general, there may not exist a joint distribution yielding a pair of families where each is the conditional pdf of the other; that is, a conditional specification of the joint distribution may not be compatible without further checking. In the present case, the Inverse Bayes Formula (1.3) yields a proper prior pdf. Hence, $L(y|\theta)$ and $p(\theta|y)$ are compatible. Indeed, by (1.3) we have

$$\pi(\theta) = \left\{e^{|\theta|}\int_0^\infty e^{-y/2}dy\right\}^{-1} = e^{-|\theta|}/2,$$

which is the standard Laplace (or double exponential) distribution, confirming the compatibility. Note that the compatibility can also be confirmed by the successful factorization

$$\begin{aligned} L(y|\theta)\times\pi(\theta) &= 2^{-1}(2\pi y)^{-1/2}\exp(-\theta^2/2y - y/2) \\ &= (2\pi y)^{-1/2}\exp(-\theta^2/2y)\times 2^{-1}e^{-y/2} \\ &= p(\theta|y)\times g(y), \quad -\infty < \theta < \infty,\ y > 0, \end{aligned}$$

where the unconditional pdf of data, $g(y)$, is an exponential distribution with mean 2.

The prior distribution so derived from the posterior distribution concentrates more around the origin than a normal prior. Here we can envisage a particle performing a linear Brownian motion on the surface of some medium for which the prior expectation of the velocity (ignoring direction, i.e. $|\theta|$) of the particle is as close to zero as in an exponential distribution.

Now let us find BIG_π and Γ_π, according to (2.4) and (2.5). In general, the repeated integral can be calculated by numerical integration methods. For the present case, the order of integration can be interchanged and we can find the exact solution for the double integral

$$\iint L(y|\theta)p(\theta|y)dyd\theta =$$

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} e^{|\theta|}(2\pi y)^{-1}\exp(-\theta^2/y - y/2)\ dyd\theta.$$

First, note that the integrand as a function of θ is symmetric about zero. The double integral equals twice the double integral that is restricted to positive θ,

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} e^{|\theta|}(2\pi y)^{-1}\exp(-\theta^2/y - y/2)\ dyd\theta =$$

$$\int_{0}^{\infty}\int_{0}^{\infty} e^{\theta}(\pi y)^{-1}\exp(-\theta^2/y - y/2)\ dyd\theta.$$

Changing variables (θ, y) to (t, s) by $\theta = ts$ and $y = s$ with the Jacobian being equal to s, we have

$$\int_{0}^{\infty}\int_{0}^{\infty} e^{\theta}(\pi y)^{-1}\exp(-\theta^2/y - y/2)\ dyd\theta =$$

$$\pi^{-1}\int_{0}^{\infty}\left\{\int_{0}^{\infty}\exp\{-s(t^2 - t + 1/2)\}\ ds\right\}dt = 3/2.$$

Hence we obtain the following from (2.4) and (2.5),

$$BIG_\pi = 3/2 - 1 = 1/2, \quad \text{and} \quad \Gamma_\pi = 1 - (3/2)^{-1} = 1/3.$$

In real applications, numerical computation is usually employed. Suppose we can sample from $\pi(\theta)$ and from $L(y|\theta)$ for a given θ; i.e. the joint distribution of (Θ, Y) can thus be sampled. For the $BIGD_\pi(y)$ in (2.6) with a particular y, we can compute $f(y)$ by Monte Carlo integration of $L(y|\theta)$ through sampling from the given prior $\pi(\theta)$,

$$f(y) \approx \sum_{i=1}^{M} L(y|\theta_i)/M, \quad \theta_i \sim \pi(\theta). \tag{2.10}$$

Then through sampling from $p(\theta|y) = \pi(\theta)L(y|\theta)/f(y)$ with the computed $f(y)$, we can obtain the conditional expectation given y by Monte Carlo again,

$$E_{p(\theta|y)}[L(y|\theta)] \approx \sum_{i=1}^{M} L(y|\theta_i)/M, \quad \theta_i \sim p(\theta|y). \tag{2.11}$$

For cases where the propagation of computational error by a numerical $f(y)$ is of concern, we can instead sample from $p(\theta|y)$ by any one of the following methods that do not not require the normalizing constant $1/f(y)$, namely the Rejection Method of von Neumann (1951), Adaptive Rejection Sampling (Gilks & Wild, 1992), Metropolis Sampling (Metropolis *et al*, 1949, 1953), Metropolis-Hastings Sampling (Hastings, 1970), the SIR method (Rubin, 1987 and 1988), and others. The MCMC methods and Gibbs sampling may be used in conjunction with the above variate-generating methods in the above sampling processes if stationarity can be assured (or trusted) on termination of iterations; see Gelman *et al.* (2004).

In regard to the information gain $BIGP_\pi(\theta)$ in (2.1) for a particular θ, we can compute $E_{L(y|\theta)}[p(\theta|y)]$ in the formula by drawing a sample $(y_1, \cdots, y_{M_2})$ from $L(y|\theta)$ and taking average of $p(\theta|y_j)$. Since each $f(y_j)$ in the expression $p(\theta|y_j) = \pi(\theta)L(y_j|\theta)/f(y_j)$ is computed as in (2.10) with a sample from $\pi(\theta)$, say a common sample $(\theta_1, \cdots, \theta_{M_1})$

for all $f(y_j)$, we therefore have

$$E_{L(y|\theta)}[p(\theta|y)] \;\approx\; \pi(\theta)\,\frac{M_1}{M_2}\,\sum_{j=1}^{M_2}\{L(y_j|\theta)/\sum_{i=1}^{M_1} L(y_j|\theta_i)\}, \quad (2.12)$$

where $\theta_i \sim \pi(\theta), \quad y_j \sim L(y|\theta)$.

Now for the total Bayesian Information Gain, BIG_π, the numerical computation for the repeated integral in (2.4) or (2.8) needs caution, as the second integration may lead to infinity. Assume, however, the repeated integral in (2.4) is finite. Then by $p(\theta|y) = \pi(\theta)L(y|\theta)/f(y)$, we have

$$\int_{\mathcal{S}(\Theta)} \left\{E_{L(y|\theta)}[p(\theta|y)]\right\} d\theta = \int_{\mathcal{S}(\Theta)} \left\{\int_{\mathcal{S}(Y|\theta)} \frac{L^2(y|\theta)}{f(y)} dy\right\} \pi(\theta) d\theta. \quad (2.13)$$

This integral can be interpreted as the expectation of the function $L(Y|\Theta)/f(Y)$ with respect to the joint density $\pi(\theta)L(y|\theta)$ for (Θ, Y). When it is finite, we can draw a sample $\{(\theta_1, y_1), \cdots, (\theta_M, y_M)\}$ from $\pi(\theta)L(y|\theta)$ by first generating θ_i from $\pi(\theta)$ and then y_i from $L(y|\theta_i)$, $i = 1, \ldots, M$. Applying (2.10) to each $f(y_k)$ with the same sample $\{\theta_i\}$ from $\pi(\theta)$ before plugging in $L(Y|\Theta)/f(Y)$, we have

$$\int_{\mathcal{S}(\Theta)} \left\{E_{L(y|\theta)}[p(\theta|y)]\right\} d\theta \;\approx\; \sum_{k=1}^{M}\{L(y_k|\theta_k)/\sum_{i=1}^{M} L(y_k|\theta_i)\}, \quad (2.14)$$

where $\theta_k \sim \pi(\theta),\ y_k \sim L(y|\theta_k)$.

3. Pairwise dependence measure and Pearson's ϕ^2

As we have seen, the concepts of Bayesian information gain function and total information gain after the Bayesian processing arise naturally. We have shown that the stronger the dependence between the data and the parameter, the more information we can gain. In this section, we discuss how these concepts, when extended to the general distributional set-up, can be employed to measure pairwise dependence between two random variables, or two random vectors.

Let X and Y be a pair of random variables, or random vectors. We use the following notation that is more common in distribution theory: $f_{X|Y}(x|y)$, $f_{Y|X}(y|x)$, $f_X(x)$, $f_Y(y)$, and $f_{XY}(x,y)$ shall denote either the pdf (probability density function) or the pmf (probability mass function) depending on whether the distribution indicated in the subscript is continuous or discrete. By the same token, $Z = f_{Y|X}(y|X)$ is a random variable as a function of X through its being the second argument of the pdf of the conditional distribution of Y given X, and $E_{X|Y=y}[f_{Y|X}(y|X)]$ denotes the expectation of Z with respect to the conditional distribution of X given $Y = y$.

The X here plays the same mathematical role of Θ as in the previous section, but is symmetrical in its relationship with Y. In view of the total Bayesian gain in (2.4) and the equal footing of X and Y, we define $\psi^2(X,Y)$, the *Pairwise Dependence Measure*, between X and Y as:

$$\begin{aligned}
\psi^2(X,Y) &\equiv \int E_{X|Y=y}[f_{Y|X}(y|X)]dy - 1 \qquad (3.1)\\
&= \int \left\{ \int f_{X|Y}(x|y) f_{Y|X}(y|x) dx \right\} dy - 1 \\
&= \int E_{Y|X=x}[f_{X|Y}(x|Y)]dx - 1 \qquad (3.2)\\
&= \int \left\{ \int f_{X|Y}(x|y) f_{Y|X}(y|x) dy \right\} dx - 1,
\end{aligned}$$

provided that the two repeated integrals in the above are equal. Here and in the sequel, we shall omit the specification of various supports of the random variables or vectors in the integrals for simplicity. Note that $\psi^2(X,Y) \equiv \psi^2(Y,X)$ and we simplify the notation to ψ^2. Now, corresponding to the Information Gain Index (2.5), we define the *Pairwise Dependence Index*, δ, by

$$\delta \equiv \psi^2/(1+\psi^2) = 1/(1+1/\psi^2) \qquad (3.3)$$

which takes values in the range $0 \le \delta \le 1$. Note that $\delta = 0$ if and only if X and Y are independent. The example below shows that for

bivariate standard normal distribution with correlation coefficient ρ, we have $\delta = \rho^2$, suggesting a benchmark comparison of δ with ρ^2.

Example 4: Consider the bivariate normal distribution with zero means, unit variances and correlation coefficient ρ. Although the bivariate pdf exists in this case, we shall use only the conditional pdf's in order to keep to the spirit of our conditional approach. Now, $Y|X = x \sim N(\rho x, 1-\rho^2)$ and $X|Y = y \sim N(\rho y, 1-\rho^2)$. Thus, the repeated integral in (3.1) is the same as that in (3.2) due to the symmetry in x and y, and is given by

$$I \equiv \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \frac{1}{2\pi(1-\rho^2)} \exp\left\{ -\frac{(y-\rho x)^2 + (x-\rho y)^2}{2(1-\rho^2)} \right\} dx \right\} dy.$$

The exponential term in the integrand can be re-written as the product:

$$\exp\left\{ -(1/2B)(x-A)^2 \right\} \exp\{-(B/2)y^2\},$$

where $A = 2\rho y/(1+\rho^2)$ and $B = (1-\rho^2)/(1+\rho^2)$. Thus,

$$I = \int_{-\infty}^{\infty} f_1(x|y)dx \times \int_{-\infty}^{\infty} f_2(y)dy \times \frac{1}{1-\rho^2},$$

where $f_1(x|y)$ is the pdf of $N(A, B)$ and $f_2(y)$ is the pdf of $N(0, B^{-1})$. This implies that $I = 1/(1-\rho^2)$, so the pairwise dependence index is, by (3.3),

$$\delta = 1 - 1/I = 1 - (1-\rho^2) = \rho^2.$$

Example 2 (Continued):

In Example 2, let θ and y be X and Y, $\delta = \Gamma_\pi$ and $\rho^2 = 1/(1+\lambda^2)$. The following table compares the two quantities for various values of λ:

λ	1	2	3	4	5	6	7	8	9	10
δ	1	0.392	0.279	0.217	0.178	0.151	0.131	0.116	0.104	0.094
ρ^2	0.500	0.200	0.100	0.059	0.038	0.027	0.020	0.015	0.012	0.010

Example 3 (Continued) In Example 3, $\Gamma_\pi = 1/3$. Now, let θ be the X variable. The pairwise dependence index between X and Y is $\delta = \Gamma_\pi = 1/3$. It turns out that the squared correlation coefficient $\rho^2 = 0$, clearly not as useful as a measure of dependence. To see this, the joint pdf is

$$f_{XY}(x,y) = \frac{1}{2\sqrt{2\pi\, y}}\, e^{-\frac{1}{2}(x^2 y^{-1}+y)},\ y>0,\ -\infty < x < \infty.$$

It can be shown that $E(XY) = 0$, by means of the bivariate transformation of $s = \sqrt{y}$ and $t = x/y$ in the double integral. We omit the detail.

When the joint distribution of two or more random variables (or random vectors), say (X, Y, Z), does not degenerate into a lower dimensional subspace, so that the joint pdf $f_{XYZ}(x,y,z))$ is defined, Pearson's ϕ^2 (1904) for measuring multi-party dependence is defined as

$$\phi^2 = \iiint \frac{f^2_{XYZ}(x,y,z)}{f_X(x) f_Y(y) f_Z(z)}\, dxdydz - 1. \tag{3.4}$$

See Joe (1989, p.161) for a discussion of ϕ^2 and for more references. In the case of a pair of random variables (or random vectors), we have

$$\begin{aligned} \phi^2 &= \iint \frac{f^2_{XY}(x,y)}{f_X(x) f_Y(y)}\, dxdy - 1 \\ &= \iint f_{X|Y}(x|y) f_{Y|X}(y|x)\, dxdy - 1 \\ &= \psi^2. \end{aligned} \tag{3.5}$$

Pearson's ϕ^2 uses the expectation of the ratio of the joint pdf to the product of marginal pdf's, which should be equal to unity in the case of independence, to measure the departure from independence. However, his approach cannot accommodate functional dependence of continuous random variables, which is a stronger form than statistical dependence. On the other hand, as we have seen that the ψ^2 has the advantage of being capable of handling both types of dependence, namely functional and statistical. Because its development is based on conditional expectations, its extension to more than two parties, however, is not as readily available as for ϕ^2. The following example underlines the difference between the two approaches.

Example 5: Let X have a continuous distribution on the real line, which is symmetric about zero such as $N(0,1)$, and let $Y = X^2$. It is a classic example to illustrate the limitations of ρ as a measure of association, because $\rho^2 = 0$ in this case while X and Y are obviously and strongly associated. This is in stark contrast with the measure δ, which takes the value 1 in this case, thus achieving the maximum of its range as shown below. Owing to the deterministic relationship, the joint distribution of (X, Y) is degenerate and restricted to the parabola, $Y = X^2$, in the upper-half of the $X \times Y$ plan, and does not have a joint pdf $f_{XY}(x, y)$ with respect to the familiar Lebesgue measure on the plan. Therefore, ϕ^2 is not defined. However, the induced conditional distributions are discrete distributions with genuine probability mass functions on the parabola. The conditional pmf $f_{Y|X}(y|x)$ is an atom along the parabola $y = x^2$,

$$f_{Y|X}(y|x) = 1 \text{ if } y = x^2, \text{ and 0 otherwise.}$$

The other conditional pmf $f_{X|Y}(x|y)$ is defined as:

When $y > 0 : f_{X|Y}(x|y) \;=\; 1/2$ if $x = -\sqrt{y}$ or $\sqrt{y}$, and 0 otherwise.

When $y = 0 : f_{X|Y}(x|y) \ = \ 1$ if $x = 0$, and 0 otherwise.

So we have $E_{X|Y=0}[f_{Y|X}(0|X)] = 1$ and for $y > 0$,

$$E_{X|Y=y}[f_{Y|X}(y|X)] = 1/2 \times 1 + 1/2 \times 1 = 1.$$

so that

$$\int_{-\infty}^{\infty} E_{X|Y=y}[f_{Y|X}(y|X)]dy = \infty.$$

Similarly,

$E_{Y|X=x}[f_{X|Y}(x|Y)] = 1/2$ for all x, except that $E_{Y|X=0}[f_{X|Y}(0|Y)] = 1$ and

$$\int_{-\infty}^{\infty} E_{Y|X=x}[f_{X|Y}(x|Y)]dx = \infty.$$

Hence $\psi^2 = \infty$ and $\delta = 1$, reflecting the deterministic relationship between X and Y.

4. DISCUSSIONS

Inspired by the essential ideas behind Fisher's information function from a frequentist framework to a Bayesian framework, we have proposed a natural measure, denoted by $BIGP_{\pi}(\theta)$, of the average information gained at any particular value of θ when the data collection set-up is repeated indefinitely. Under positivity condition, this gain actually equals the arithmetic mean minus the harmonic mean of the posterior distribution with respect to the likelihood. Integrating the function $BIGP_{\pi}(\theta)$ with respect to θ we have the total Bayesian information gain BIG_{π} and an index Γ_{π} taking value between 0 and 1 as a normalized measure. In regard to Bayesian data analysis, we also consider a measure of Bayesian information gain conditional to datum at hand, denoted by $BIGD_{\pi}(y)$, and its normalized index, $\Gamma_{\pi}(y)$. We have used very simple examples for focused demonstration

on the ideas and then indicated the way forward for numerical computation. Reversing the Bayes' formula is a long-neglected thought process by Bayesian statisticians. Our examples, especially Example 3, have highlighted the fact that the inverse Bayes' formula can lead to many unexpected consequences.

Noting that the amount of information gain measures the degree of dependence between the data and parameter, we have extended the measure to the general distribution theory, in which the data and parameter are treated, on equal footing, as a pair of random variables, or random vectors. We are thus led to the measure, ψ^2, for pairwise dependence by defining it in terms of the two relevant conditional distributions. When the unconditional joint pdf is defined, ψ^2 reduces to Pearson's ϕ^2. However, one advantage enjoyed by ψ^2 is that it can handle functional dependence while ϕ^2 cannot. We have also introduced the Pairwise Dependence Index, denoted by δ, which is capable of revealing non-linear association. Moreover, δ can achieve its maximum of one when the variables are functionally related, thus completing the spectrum from dependence to independence of two random variables or two random vectors.

Of course, $(p(\theta|y) - \pi(\theta))$ is not the only possible sensitivity function. An alternative is $\log(p(\theta|y)/\pi(\theta))$, for which we have the following *Bayesian Information Gain on Parameter in Log*:

$$BIGPL_{\pi}(\theta) \equiv E_{L(y|\theta)} \log(p(\theta|y)/\pi(\theta)) = E_{L(y|\theta)} \log p(\theta|y) - \log \pi(\theta). \tag{4.1}$$

Its similarity to the Kullback-Leibler measure permits similar interpretation. The requirement for $BIGPL_{\pi}(\theta) \geq 0$ is guaranteed by the following inequalities

$$\exp\left\{\int_{\mathcal{S}(Y|\theta)} L(y|\theta) \log p(\theta|y) dy\right\} \geq \left\{\int_{\mathcal{S}(Y|\theta)} \frac{L(y|\theta)}{p(\theta|y)} dy\right\}^{-1} \geq \pi(\theta), \tag{4.2}$$

where the first inequality is between the geometric mean on the left-hand side and the harmonic mean on the right-hand side, while the second is just (1.2). Analogous to (2.4) and (2.5), we have the *Bayesian Information Gain in log*,

$$BIGL_\pi \equiv \int_{\mathcal{S}(\Theta)} BIGPL_\pi(\theta)d\theta \tag{4.3}$$

$$= \int_{\mathcal{S}(\Theta)} \left\{ \int_{\mathcal{S}(Y|\theta)} L(y|\theta) \log[p(\theta|y)/\pi(\theta)]dy \right\} d\theta, \tag{4.4}$$

and the *Information Gain Index on log scale*, or $Log\Gamma_\pi$ for short,

$$Log\Gamma_\pi \equiv BIGL_\pi/(1 + BIGL_\pi). \tag{4.5}$$

Acknowledgments

The authors thank the referee and the Associate Editor for their comments. They are grateful to the referee for his pertinent questions and constructive suggestions, which have helped removing errors and obscurities in the computation. The joint research was supported by HKU small-projects funding. It was conducted during HT's visit to The University of Hong Kong as a distinguished visiting professor and the paper was completed during his tenure as the Saw Swee Hock Professor at the National University of Singapore.

References

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*, Second Edition, London: Chapman & Hall.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, **57**, 97-109.

Joe, H.(1989). Relative Entropy Measures of Multivariate Dependence. *J. Amer. Stat. Ass.*, **84**, 157-164.

Johnson, N.L., Kotz, S. and Balakrishan, N. (1994). *Continuous Univariate Distributions Vol 1*, 2nd Ed. New York: John Wiley & Sons.

Meng, X.L. (1996). Comments on "Statistical inference and Monte Carlo algorithms" by G. Casella. *Test* **5**(2), 310-318.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo Method. *J. Amer. Stat. Ass.*, **44**, 335-341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **21**, 1087-1092.

Ng, K.W. (1995). Explicit formulas for unconditional PDF (Rev. March 1995). Research Report No. 82, Department of Statistics and Actuarial Science, The University of Hong Kong.

Ng, K.W. (1997). Inversion of Bayes Formula: Explicit Formulae for Unconditional pdf, *Advances in the Theory and Practice of Statistics* (edited by Norman L. Johnson and N. Balakrishnan) Chapter 37, pp. 571-584. New York: John Wiley and Sons.

Pearson, K. (1904). Mathematical Contributions to the Theory of Evolution, XIII: On the Theory of Contingency and Its Relation to Association and Normal Correlation, in *Drapers Company Research Memories* (Biometric Series I). London: University College [reprinted in *Early Statistical Papers* (1984) by the Cambridge University Press,

Cambridge, U.K.].

Rubin, D.B. (1987). Comments on "The calculation of posterior distributions by data augmentation" M.A. Tanner & W.H. Wong. *J. Am. Statist. Assoc.* **82**, 543-546.

Rubin, D.B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussions). In *Bayesian Statistics, Vol.3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, eds.), 395-402. Oxford University Press, Oxford.

Tan, M., Tian, G. and Ng, K.W. (2009) *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*, London: Chapman & Hall/CRC.

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd Ed., New York: Springer.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Assoc.* **82**, 528-540.