THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #255

UNCERTAINTY AND THE CONDITIONAL VARIANCE

BY

JIAHUA CHEN
CONSTANCE VAN EEDEN
JAMES V ZIDEK

January 2010

Uncertainty and the Conditional Variance

By Jiahua Chen, Constance van Eeden and James Zidek
Department of Statistics, University of British Columbia, Vancouver,
British Columbia, V6T 1Z2 Canada
jhchen@stat.ubc.ca; vaneeden@stat.ubc.ca; jim@stat.ubc.ca.

Summary

Statisticians have long viewed the quest for more information, for example through the acquisition of additional data, as being central to the goal of reducing uncertainty about some aspect of the world. This paper explores that objective through the variance, a common way of quantifying uncertainty. In particular, it examines the relationship between information and uncertainty. Surprisingly it shows that increasing the amount of information can in some cases increase the variance while in others it can decrease it. Which of these occurs is not explained by the seductive thesis that it depends simply on whether that uncertainty is merely aleatory - due to chance alone - or epistemic - due to lack of knowledge. Through examples it shows the relationship to be complex and a general theory elusive.

Keywords: Information; Normal distribution; Uncertainty; Variance.

1 Introduction

This paper continues an investigation begun in Zidek and van Eeden (2003; hereafter ZvE), of the relationship between information and uncertainty. Our focus is on the effect of increasing information on the prediction of a random object X. That object could be a random variable, a parameter in the Bayesian framework adopted in this paper, or an estimator. Our measure of uncertainty is the variance of X i.e. $\text{VAR}(X)$, chosen both for its simplicity as well as its popularity in statistical science where for example the standard error is commonly used to index estimator uncertainty. In particular, we present the solution to a problem posed by ZvE that has remained open until now: “Is the conditional variance $\text{VAR}(X; |X| < c)$ an increasing function of c as one intuitively expect, when $X \sim N(\eta, 1)$ for some parameter η?”

The importance of uncertainty transcends its fundamental role in statistical science and in fact, DeFinetti says

The only relevant thing is uncertainty - the extent of our own knowledge and ignorance. (de Finetti, 1970/1974, preface, xi-xii)

Both Google and Google Scholar, yield a huge number of sites concerned with uncertainty or its cousin "information". Yet despite its importance, it is not very well–defined, and in that respect ZvE liken it to its cousin about which Basu (1975) said:

But, what is information? No other concept in statistics is more elusive in its meaning and less amenable to a generally agreed definition.

Whatever its definition, the need to quantify uncertainty seems uncontroversial, as that concept is commonly used in a comparative sense in ordinary communication where phrases such as "greater uncertainty" and "more uncertain" are commonly encountered. ZvE discuss ways of doing so. One such approach is through probability as in Frey and Rhodes (1996) who propose the use of the probability distribution for that purpose. O'Hagan (1988) uses probability as measure of "uncertainty".

However, although probability provides a framework in which to discuss uncertainty, it does not provide the index of uncertainty we need to make comparative statements of the type referred to above. Thus Harris (1982) proposes the (relative) entropy of a probability distribution as such an index. In fact that notion of entropy goes back at least as far as Shannon (1948) when uncertainty and information were seen as identical; the quantitative uncertainty $U(X)$ about a randomly distributed object X is thought of as the amount of information observing X would provide, since then all uncertainty about it would vanish.

While entropy is appealing, variance is simpler. Hence it is more widely used. Thus a random variable with a large variance like one with large entropy, is one with an elevated volatility, making it difficult to predict. This is the index used in this paper to explore the impact of information on uncertainty.

One might naively expect that uncertainty would be reduced when some additional information about X becomes available. More specifically, we might think that the conditional variance $g(c) = \text{VAR}(X; |X| < c)$ should decrease as c decreases since the latter means increasingly more information is being provided about X. Surprisingly, examples in Section 2 involving both discrete and continuous cases, show that this will not always be the case. In that section we show why. Briefly this phenomenon is seen when that information contradicts some aspect of a model that is confidently believed to generate X. In fact, this insight suggests that such information would always lead to an increase in uncertainty.

But this would be another failed conjecture. Surprisingly that is not always true. While the counter examples in Section 2 show that $g(c)$ is not generally an increasing function of

c, the claim is true for the family of normal distributions, as shown in Section 3. For them, we show that $g(c)$ in an increasing function of c when X has normal distribution with any mean and variance.

The next section studies in some detail two complementary types of uncertainty and the tradeoff between uncertainty and information with respect to each. However that tradeoff is a complex one and the subject of our concluding section. There we highlight some issues for future work.

2 Epistemic and aleatory uncertainty

This section examines important aspects of the uncertainty-information relationship through a simple dichotomy of the forms of uncertainty. The first, epistemic uncertainty derives from a lack of knowledge. It includes model as well as parameter uncertainty. Moreover the amount of epistemic uncertainty changes as information is acquired and knowledge developed. The second type, aleatory uncertainty is due to chance. In contrast to epistemic uncertainty, aleatory uncertainty does not change in the light of any new information that could realistically be acquired.

Within a Bayesian framework, variance captures both types of uncertainty. To see that, let θ stand ambiguously for both the model as well as its parameters. Then a familiar identity tells us that the overall uncertainty is given by

$$U(X) \equiv \text{VAR}(X) = E[\text{VAR}(X;\theta)] + \text{VAR}[E(X;\theta)]. \tag{1}$$

The first term is the aleatory uncertainty averaged over all possibilities for the uncertain θ. The second assesses the epistemic uncertainty through the model's ability to predict X. We see great economy in this identity, in that the second term re-uses the variance in a another role, the assessment of uncertainty in the predictor.

We assume that information comes from data that are thought to inform us about X, more specifically from the observation of a random object $Z = z$. The posterior uncertainty would now become $U(X;z) \equiv \text{VAR}(X; Z = z)$ once z has been observed and this paper investigates whether knowing that $Z = z$ has reduced our uncertainty.

The following example illustrates these ideas.

EXAMPLE 1 In an experiment a subject's uncertainty about the event $X = 1$ that a die toss yields an 'ace' is assessed. That subject, convinced that the toss is fair, gives probability $\theta = 1/6$ to the event $\{X = 1\}$. Using variance to express that subject's uncertainty, we find $\mathrm{VAR}(X) = \mathrm{VAR}(X; p = 1/6) = 5/36$, all aleatory uncertainty, the second term of the Equation (1) being zero.

A second subject, recalling Lindley's version of Cromwell's rule, is unwilling to assume the toss is fair. That subject puts a uniform distribution on θ to reflect complete ignorance about its value. Equation (1) now gives the uncertainty as $\mathrm{VAR}(X) = E[\mathrm{VAR}(X; \theta)] + \mathrm{VAR}[E(X; \theta)]$. Here $E[\mathrm{VAR}(X; \theta)] = E[\theta(1 - \theta)] = 1/2 - 1/3 = 6/36$, it being the expected aleatory uncertainty and higher than the first subject's 5/36. At the same time, $\mathrm{VAR}[E(X; \theta)] = \mathrm{VAR}(\theta) = 1/3 - (1/2)^2 = 3/36$, the epistemic uncertainty that was 0 for the first subject. In sum Subject 2's uncertainty is quantified as 9/36 in contrast to Subject 1's 5/36.

Now new information arrives in the form of a list of the outcomes of a sequence of n die tosses that yielded r occurrences of the event $\{X = 1\}$. Unlike the first subject's uncertainty, the second's must change as the prior is transformed into the posterior beta distribution with density proportional to $\theta^r(1-\theta)^{(n-r)}$. In fact, regarding both E and VAR as conditional on the new information,

$$E[\mathrm{VAR}(X; \theta)] = E[\theta(1 - \theta)] = \frac{(r + 1)(n - r + 1)}{(n + 3)(n + 2)} \approx \frac{r}{n}(1 - \frac{r}{n})$$

and

$$\mathrm{VAR}[E(X; \theta)] = \mathrm{VAR}(\theta) = \frac{(r + 1)(n - r + 1)}{(n + 2)^2(n + 3)}.$$

Thus the epistemic uncertainty approaches 0 as would be expected while if indeed the tosses are fair, the expected aleatory uncertainty approaches that of the first subject. ■

REMARK 1 This example demonstrates the impact information can have through changes on the epistemic uncertainty. A $\hat{\theta}$ substantially different from 1/6 would erode Subject 1's certainty in his or her model and force change. Subject 2 would be in the same position if it turned out that application of Cromwell's rule had not been aggressive enough. For example, the appearance of an 11 anywhere in the sequence of tosses would yield a 0 likelihood under any one of Subject 2's classes of models and rule out use of Bayes rule for updating the prior. In fact such an outcome leads to so–called "deep uncertainty" that represents amongst other things, the "unknown unknowns" envisioned by the former US Secretary of Defense, Donald Rumsfeld. ■

The following two examples continue this theme and show that uncertainty can increase due to more subtle effects of data than those seen in the previous example. Moreover, they show these effects can obtain in the case of both discrete and continuous random variables

X. In both cases $Z = I\{|X| < c\}$ and interest focuses on the effect of increasing information by reducing c.

EXAMPLE 2 Suppose X has probability mass function:

$$P(X = 0) = q/2; \quad P(X = 1) = q/2, \quad P(X = 2) = p$$

where $p \in (0, 1)$ and $q = 1 - p$. When $c = 1.5$, the conditional distribution is binomial with parameters (1, 0.5). Thus,

$$\text{VAR}(X; |X| \leq 1.5) = 0.25.$$

When $c = 2.5$, the condition $|X| < c$ no longer constrains. We find

$$\text{VAR}(X) \leq E(X - 2)^2 = 5q/2.$$

Thus, when q is sufficiently small,

$$g(1.5) > g(2.5)$$

which implies that $g(c)$ is not a decreasing function in c in general. ■

In the next example

$$\text{VAR}(X; |X| \leq c_1) < \text{VAR}(X; |X| \leq c_2) \text{ for } c_1 > c_2. \tag{2}$$

It extends the previous example for discrete distributions to continuous ones.

EXAMPLE 3 Let $f(x)$, $x \geq 0$ be given by

$$f(x) = \begin{cases} p & \text{for } x \in (0, 1) \\ h & \text{for } x \in (2, 2 + \varepsilon), \end{cases}$$

for some $p \in (0, 1)$, $\varepsilon \geq 0$, $h > 0$ and $h\varepsilon = 1 - p$.

Then, for $c = 1.5$, $\text{VAR}(X; |X| \leq c) = 1/12$ and for $c \geq 2 + \varepsilon$,

$$\text{VAR}(X; |X| \leq c) = \text{VAR}(X)$$

and there exist (p, ε) such that $\text{VAR}(X) < 1/12$.

This can be seen as follows:

$$E(X) = \frac{p}{2} + (1 - p)(2 + \frac{\varepsilon}{2}),$$

$$E(X^2) = \frac{p}{3} + \frac{1-p}{3}(12 + 6\varepsilon + \varepsilon^2),$$

$$\text{VAR}(X) = \frac{p}{3} + \frac{1-p}{3}(12 + 6\varepsilon + \varepsilon^2) - \Big(\frac{p}{2} + (1-p)(2 + \frac{\varepsilon}{2})\Big)^2.$$

Note that when $p = 0$,

$$\text{VAR}(X) = \frac{\varepsilon^2}{12} < \frac{1}{12}$$

for every $\varepsilon \in (0, 1)$. Because $\text{VAR}(X)$ is continuous in p, for every $\varepsilon \in (0, 1)$, there exists an interval of small p-values for which the condition

$$\text{VAR}(X; |X| \leq c) < \frac{1}{12} = \text{VAR}(X; |X| \leq 1.5) \text{ when } c > 2 + \varepsilon$$

is satisfied. ■

REMARK 2 In the foregoing example, the shapes of the two parts of the density do not have to be uniform. What is needed is that the lefthand part carries a small part of the total mass and has a large conditional variance while the righthand side is a large part of the total mass and has a small conditional variance. ■

This section has demonstrated the benefits of the Bayesian framework for studying uncertainty since it accommodates both aleatory and epistemic uncertainty. In fact there is no need to distinguish between them and this feature can be exploited to achieve parsimony and hence computational efficiency in constructing and implementing process models.

However, ignoring the distinction can also lead to difficulties since the epistemic component of uncertainty may well be reduced or even disappear as information comes in. The result can be models which fail to preserve the process's marginal aleatory uncertainty. The following example illustrates the problem.

EXAMPLE 4 A process $\{X_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T\}$ is measured over time t at a number of sites p and the goal is a model for this space-time process. The analyst chooses the following one:

$$\begin{aligned}
X_{it}|\beta &= \beta_{it} + \epsilon_{it},\ i = 0, 1, \ldots, p,\ t = 1, \ldots, T+1, \text{ or} \\
\mathbf{X}_i|\beta &= \beta_i + \epsilon_i,\ i = 0, 1, \ldots, p, \text{ with} \\
\beta_i &\sim \mathrm{N}(\mu_0 1_{T+1}, \sigma^2_\beta\ \mathrm{I}_{T+1}),\ i = 0, 1, \ldots, p \\
\mu_0 &\sim N(\mu^*, \sigma^2_\mu),
\end{aligned}$$

where the variances are known while the ϵ's are independent of one another. For parsimony, the model thus puts all the spatial correlation into that of the β's. Some of the randomness in the β's might be due to chance, such things as fluctuations in temperature or wind direction. This randomness and hence uncertainty about the coefficients would thus be aleatory. But some of their uncertainty would be due to their unknown baseline level μ_0 and would be epistemic.

However the parsimony comes at considerable cost. For as data are acquired, the epistemic uncertainty is lost and a posteriori, the processes become spatially uncorrelated. More precisely if we let $\mathrm{X} = (\mathrm{X}_1', \ldots, \mathrm{X}_T')' : pT \times 1$ denote the column vector of observed responses to time T, the spatial covariance between sites i and j at time $T+1$ is

$$\begin{aligned}\textsc{cov}(\mathrm{X}_{i(T+1)}, \mathrm{X}_{j(T+1)}; X) &= \textsc{var}(\beta; \mathrm{X}) \\ &= \sigma_\mu^2 - \sigma_\mu^2 1'[\sigma^2 \mathrm{I} + \sigma_\mu^2 \mathrm{J}]^{-1} 1 \sigma_\mu^2\end{aligned}$$

where 1 stands for the pT vector all of whose element are 1, I the corresponding identity matrix, $\mathrm{J} = 11'$ and $\sigma^2 = \sigma_\beta^2 + \sigma_\epsilon^2$. Using the familiar matrix identity, $[\mathrm{A} + \mathrm{B}'\mathrm{C}\mathrm{B}]^{-1} = \mathrm{A}^{-1} - \mathrm{A}^{-1}\mathrm{B}'[\mathrm{C}^{-1} + \mathrm{B}\mathrm{A}^{-1}\mathrm{B}']^{-1}\mathrm{B}\mathrm{A}^{-1}$ the previous equation implies

$$\textsc{cov}(\mathrm{X}_{i(T+1)}, \mathrm{X}_{j(T+1)} | X) = [\sigma_\mu^{-2} + pT\sigma^{-2}]^{-1} \to 0, \text{ as } T \to \infty.$$

As a representation of the analyst's belief, this last result would be completely unrealistic. ■

REMARK 3 Bayesian models used to represent space–time processes are substantially more sophisticated than the simple one in the example. Nevertheless the loss of epistemic uncertainty will occur over time and the impact of that losses needs to assessed as a standard diagnostic test of the model. ■

Section 4 will discuss the lessons learned from the examples in this section.

3 The main result

The following theorem contains the main technical result of this paper.

THEOREM 1 *Let X be a normally distributed random variable with mean θ and variance σ^2. Then the conditional variance,* $\textsc{var}(X; |X| \le c)$ *is an increasing function of c.* ■

When $\theta = 0$, the result is easy to prove. It has also been shown that $\textsc{var}(X; |X| \le c) < \sigma^2$ for any θ (Zidek and van Eeden, 2003). We present the proof in several steps which begin with one that symmetrizes the problem.

If $\text{VAR}(X; |X| \leq c)$ is an increasing function in c, then its derivative must be positive. However proving that fact by a straightforward approach proves difficult because the relevant quantities do not have easy–to–handle analytical forms. In particular, subtracting the square of the conditional mean makes the expression of the conditional variance messy. Instead we approach the problem by an indirect route that bypasses the need to examine the conditional mean.

The alternative route relies on identically distributed and independent random variables, Y_1 and Y_2, each having the conditional distribution of X given $|X| \leq c$. Clearly $2\text{VAR}(X; |X| \leq c) = E(Y_1 - Y_2)^2$. Thus the required monotonicity becomes that of

$$g(c) = E\{(Y_1 - Y_2)^2\}.$$

For notational simplicity, we replace the conditioning event by $|X| \leq c/\sqrt{2}$. In addition, without loss of generality, we assume $\theta > 0$ and $\sigma = 1$. Given $|X| \leq c/\sqrt{2}$, the conditional density of X is found to be

$$\frac{\phi(x-\theta)I(|x| \leq c/\sqrt{2})}{\Phi(-\theta + c/\sqrt{2}) - \Phi(-\theta - c/\sqrt{2})}$$

where $\phi(x)$ and $\Phi(x)$ are the density function and cumulative distribution function of the standard normal distribution. Let Z be a standard normally distributed random variable. The denominator can then be written as

$$P(|Z + \theta| \leq c/\sqrt{2}).$$

and the density is proportional to $\phi(x - \theta)$ within the specified range.

Let

$$U = (Y_1 - Y_2)/\sqrt{2} \text{ and } V = (Y_1 + Y_2)/\sqrt{2}.$$

This transformation has Jacobian equal to 1 and the range of (u, v) is given by

$$\{(u, v) : |v| < c - u, \quad 0 < u < c\}$$

and its mirror image with respect to the $u = 0$ axis. The joint density function of U and V is (within the range specified earlier)

$$\frac{\exp\{-\frac{1}{2}[u^2 + v^2 - 2\sqrt{2}\theta v + 2\theta^2]\}}{2\pi P^2(|Z + \theta| \leq c/\sqrt{2})}$$

and the marginal density function of U is, for $0 < u < c$ given by

$$h(u; c) = \phi(u)\frac{P(|Z + \sqrt{2}\theta| \leq c - u)}{P^2(|Z + \theta| \leq c/\sqrt{2})}.$$

This density function is, for $-c < u < 0$ given by symmetry: $h(u, c) = h(-u, c)$.

For any $0 < c_1 < c_2$, let U_1 and U_2 be two random variables with distribution $h(u; c_1)$ and $h(u; c_2)$ respectively. Our task now becomes showing that

$$g(c_1) = E\{U_1^2) < g(c_2) = E\{U_2^2\}.$$

We show this result is true in the following lemma.

LEMMA 1 *There exists a* u^* *between* 0 *and* c_1 *such that*

$$h(u^*, c_1) = h(u^*, c_2)$$

and $(u - u^*)\{h(u, c_1) - h(u, c_2)\} \leq 0$ *for any* $u \geq 0$.

REMARK 4 Before proving this lemma, it is useful to note that, when such a u^* exists, we have

$$\begin{aligned}
& E\{U_1^2\} - E\{U_2^2\} \\
= \;& 2\int_0^{u^*} u^2\{h(u; c_1) - h(u; c_2)\}du + 2\int_{u^*}^{c_2} u^2\{h(u; c_1) - h(u; c_2)\}du \\
\leq \;& 2(u^*)^2 \int_0^{u^*} \{h(u; c_1) - h(u; c_2)\}du + 2(u^*)^2 \int_{u^*}^{c_2} \{h(u; c_1) - h(u; c_2)\}du \\
= \;& 0.
\end{aligned}$$

Thus, the theorem is proved once this lemma is proved. ■

PROOF OF LEMMA 1. We first show that $\log h(u; c)$ is a decreasing function of u in the range $(0, c)$ for any given $c > 0$. This is straightforward because

$$\begin{aligned}
\log h(u, c) \;=\; & -\frac{1}{2}u^2 + \log\{\Phi(\sqrt{2}\theta + c - u) - \Phi(\sqrt{2}\theta - c + u)\} \\
& +2\log P(|Z + \theta| \leq c/\sqrt{2}) - \frac{1}{2}\log(2\pi)
\end{aligned}$$

and

$$\frac{d \log h(u, c)}{du} = -u - \frac{\phi(\sqrt{2}\theta + c - u) + \phi(\sqrt{2}\theta - c + u)}{\Phi(\sqrt{2}\theta + c - u) - \Phi(\sqrt{2}\theta - c + u)} < 0.$$

Next, we claim that the above derivative at any given u is increasing in c. This claim is equivalent to

$$\frac{\phi(\sqrt{2}\theta + c - u) + \phi(\sqrt{2}\theta - c + u)}{\Phi(\sqrt{2}\theta + c - u) - \Phi(\sqrt{2}\theta - c + u)}$$

being a decreasing function of c, i.e. to the derivative of

$$\log\{\phi(\sqrt{2}\theta + c - u) + \phi(\sqrt{2}\theta - c + u)\} - \log\{\Phi(\sqrt{2}\theta + c - u) - \Phi(\sqrt{2}\theta - c + u)\}$$

being non-positive. For simplicity, we replace $\sqrt{2}\theta$ by θ and $c - u$ by t. Then the function becomes

$$w(t) = \log\{\phi(\theta + t) + \phi(\theta - t)\} - \log\{\Phi(\theta + t) - \Phi(\theta - t)\}$$

and we show that its derivative with respect to t is non-positive taking note of the fact that both θ and t are positive.

For this goal, we compute the derivative with respect to t and find

$$\begin{aligned} w'(t) &= \frac{(\theta - t)\phi(\theta - t) - (\theta + t)\phi(\theta + t)}{\phi(\theta + t) + \phi(\theta - t)} - \frac{\phi(\theta + t) + \phi(\theta - t)}{\Phi(\theta + t) - \Phi(\theta - t)} \\ &= \frac{\theta\{\phi(\theta - t) - \phi(\theta + t)\}}{\phi(\theta + t) + \phi(\theta - t)} - t - \frac{\phi(\theta + t) + \phi(\theta - t)}{\Phi(\theta + t) - \Phi(\theta - t)} \\ &\leq (\theta - t) - \frac{\phi(\theta + t) + \phi(\theta - t)}{\Phi(\theta + t) - \Phi(\theta - t)}. \end{aligned}$$

When $\theta - t \leq 0$, $w'(t)$ is clearly negative. When $\theta - t > 0$, we have

$$\Phi(\theta + t) - \Phi(\theta - t) \leq 1 - \Phi(\theta - t) \leq \frac{\phi(\theta - t)}{(\theta - t)}$$

by the well-known inequality $1 - \Phi(x) \leq x^{-1}\phi(x)$ for any $x > 0$. This again implies that $w'(t)$ is negative.

To conclude, we have seen that for given $0 < c_1 < c_2$ and $\theta > 0$, both $\log h(u, c_1)$ and $\log h(u, c_2)$ are decreasing functions of u. In addition,

$$\frac{\partial h(u, c_1)}{\partial u} < \frac{\partial h(u, c_2)}{\partial u}$$

for any $u \in [0, c_1]$.

Because the area under two density functions are equal, they must intersect at some u^*. At the same time, because the rate of decrease of $\log h(u, c_1)$ is always larger than that of $\log h(u, c_2)$, there can be at most one intersection between $h(u, c_1)$ and $h(u, c_2)$. This proves Lemma 1 and subsequently also the main theorem. ■

4 Discussion

The effect of additional information on the uncertainty about a random object X is difficult to characterize even when that information comes in the form of an observed value of

$Z = I\{|X| < c\}$. This paper has shown the naive expectation that decreasing c, and hence increasing the amount of information about X, would decrease $\text{VAR}(X; |X| < c)$ is shown in Section 2 to be false in general. As well, we see in that section the reason why and that is because the additional information can be in conflict with some feature of the model that is thought to generate X. Thus the validity of the model is thrown into doubt. This discovery leads to the conjecture that this is true in general.

Not true. In Section 3 we see the conjecture fails for an important family, the class of normal distributions, making a general theory elusive. From a design perspective when prediction of X is the inferential objective, one might ask if observing Z is worthwhile. The answer to that is positive if cost is ignored since

$$\text{VAR}(X) - E[\text{VAR}(X; Z)] = \text{VAR}[E(X; Z)] > 0 \quad (3)$$

except when X and Z are uncorrelated. Thus observing Z should reduce uncertainty about X. Nevertheless, as the above examples show, the result could well be an increase in uncertainty, the likelihood of that happening depending on the degree of correlation between the predictor and the predictand, pointing anew to the new to select the predictor Z optimally.

To conclude we return to the case alluded to in Section 1 where uncertainty $U(X)$ and information are regarded as equal, the latter being what is gained when X is observed and all uncertainty about it is lost. Originally, the entropy in the distribution of X seems to have been thought of in that way. Moreover Fisher's information bears the same hallmark. There, if Y has density f, the point $\hat{\theta}$ at which $f(Y; \theta)$ or alternatively $\log f(Y; \theta)$ is maximized would be the θ best supported by Y. Furthermore if the random score function $X = -\partial \log f(Y; \theta)/\partial\theta$ is large at that point, Y would discriminate well between it and its local neighbors at least. Since $E(X) = 0$, a large value of Fisher's information $I(\theta) = \text{VAR}(X; \theta)$, which measures the aleatory uncertainty in the score function, says that observing X would greatly reduce uncertainty about θ. Thus uncertainty is information in this case. Curiously, the uncertainty about the maximum likelihood estimator as measured by its asymptotic variance is approximately the reciprocal of the $I(\theta) = \text{VAR}(X; \theta)$ where X is now computed from that sample.

In practice, θ would be subject to epistemic uncertainty as well, suggesting an extension of Fisher's information: $I = \text{VAR}(X) = E[I(\theta)] + \text{VAR}[E(X; \theta)] = E[I(\theta)]$. However this version of I does seem to have been proposed elsewhere.

5 Conclusions

The paper leaves us with a complex picture about the relationship between uncertainty and information. Clearly the type of information and the type of uncertainty play a role in determining that relationship. But as the variety of examples in this paper show, a general theory seems elusive and further study is needed. Furthermore, it is clear that even the simplest questions can lead to very challenging technical problems.

References

1. Basu, D. (1975). Statistical information and likelihood. *Sankhyā, Ser A*, 37, 1-71.

2. De Finetti, B. (1970/1974). *Teoria delle Probabilità* 1, Turin: Einaudi. English translation as *Theory of Probability* 1 in 1974, Chichester, Wiley.

3. Frey, H. C. and Rhodes, D. S. (1996). Characterizing, simulating, and analyzing variability and uncertainty: An illustration of methods using an air toxics emissions example. *Human and Ecological Risk Assessments.* **2** 762-797.

4. Harris, B. (1982). Entropy. In *Encyclopedia of Statistical Science* Volume **2**. Editors: S. Kotz and N.L. Johnson. Wiley, New York, 512-516.

5. O'Hagan, A. (1988). *Probability, Methods and Measurement.* Chapman and Hall, London.

6. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal.* **27**. 379-423, 623-656.

7. Zidek, J. V. and van Eeden, C. (2003). Uncertainty, entropy, variance and the effect of partial information. *"Mathematical Statistics and Applications" Festschrift for Constance van Eeden.* Lecture notes-monograph series, Institute of Mathematical Statistics. Editors: Marc Moore, Sorana Froda and Christian Leger. Volume **42**. 155-167.