International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2012), 80, 1, 83–86 doi:10.1111/j.1751-5823.2011.00169.x

Discussion

Alexandre Bouchard-Côté and James V. Zidek

University of British Columbia, Vancouver, Canada
E-mail: jim@stat.ubc.ca

The last century saw a sustained but largely failed effort to put the emerging discipline of statistics on a solid foundation. It began well with the pioneers like Fisher who introduced within a frequency theory framework, a rudimentary foundation for inference, tools for developing procedures for users and criteria such as unbiasedness by which to assess their performance. And it produced great success stories like design of experiments. Building on the success of Neyman and Pearson, Wald developed an axiomatic foundation, decision theory, that seemed intended to do for statistical science what Kolmogorov's axioms had done for probability. A program of largely mathematical research put new and old methods such as the sample average on the seemingly solid ground provided by those axioms.

But holes began to appear in it. An inexplicable and unintuitive result was published in 1956 by Stein, which cast doubt on some of its cornerstones such as unbiasedness. He showed, with an early version of the "James-Stein estimator", that bias could be traded for variance for an overall gain in overall parameter estimation precision in problems involving a large number p of different populations from which independent samples of moderate size were drawn. This puzzling result was only explained later through the notion of exchangeability within the alternate framework provided by the Bayesian paradigm which by then had emerged as a competitor with another axiomatic foundation. Not only that but Bayesians like Savage and Lindley saw irresolvable inconsistencies within the frequency theory framework.

The result was a surge in the popularity of the Bayesian approach, helped by advances in such things as statistical computation. Subject area scientists embraced it, seemingly convinced that its foundations had been assured by the statisticians. On the other hand statisticians saw its success in applications as validating it and thereby licensing them to focus on its development.

But once again holes began to appear. The celebrated work of Kahneman and Tversky cast doubt on the rationality of human decision makers and hence on the validity of the axioms that underpinned the paradigm. Meanwhile statisticians began to see difficulties with the subjective foundations and their suitability as a basis for constructing prior distributions, one of the cornerstones along with Bayes rule, of the axiomatic theory as it first appeared. So work to shore up those foundations began. Thus Sir David Cox, in his ASA Presidential Invited Address at the 2011 Joint Statistical Meetings, was able to find more than half a dozen approaches to finding those priors, revealing the morass into which the elegant and simple paradigm had sunk despite its ever-increasing popularity.

So we tucked into this paper with great relish, as it promised a foundation for a principled choice of a prior. We found that the authors have put together an elaborate and thoughtful case. But imagine our surprise when it gave us not only a prior but also as one of its main conclusions, that conventional use of Bayes rule in parametric inference is invalid! That led us to consider the key elements of their argument and our discussion below will examine them.

Their theory builds on the key idea that one should focus on the observables and their distribution. This idea is not new. A notable omission in their citations is the work of Akaike who also makes this a fundamental ingredient to his very productive approach to the theory of inference. That remarkably powerful idea yields in addition to the celebrated AIC criterion:

(i) the maximum likelihood estimator;
(ii) the superiority for prediction, of using a predictive distribution for an observable over a point estimator;
(iii) a derivation of Bayes rule in the parametric case (Akaike, 1978).

Akaike (1974) introduces his basic criterion functional as

$$-\int g(x)\log f(x|\theta)dx.$$

He regards $f(x|\theta)$ as a predictive distribution for x which has an unknown and unknowable distribution with density function g. Akaike's basic criterion is related to the one given in the paper,

$$-\int \log f(x|\theta)dF_0(x), \tag{1}$$

with the difference that the former assumes density representations for both the observable and the parametric families. Both suffer from the technical deficiency that the observable x would as a measurement have units attached as would therefore $f(x|\theta)$. This means that the logarithm cannot legitimately be evaluated at $f(x|\theta)$. This is because the equation $1+u+u^2/2+\cdots=\exp u$ means $\exp u$ and u must be unitless and hence so must $v=\exp(\log v)$ as well as $\log v$. The operational difficulties associated with equation (1) can be seen in Subsection 3.1 where the first and third terms in the expansion of the logarithm are unitless while the second $\log\lambda$ involves λ, which has units and makes this expression uninterpretable. In his subsequent work (e.g., Akaike, 1978), Akaike replaces this criterion function with

$$-\int g(x)\log[f(x|\theta)/g(x)]dx,$$

which overcomes the problem. Not only that it makes the criterion function invariant under one-to-one transformation of x as seems sensible. Jaynes (1963) emphasizes the importance of having such invariance.

It is unclear which prior should be used for the distribution F_0. The authors use a Dirichlet process, but many other choices seem possible. Consider the problem of estimating, say, tomorrow's "maximum temperature" for example. In that case, even estimating it by the empirical distribution function as the authors suggest to get the likelihood function seems unnatural. Presumably there the observable is measured on a discrete scale with round-off so that X_i would represent a value of x in an interval L_i of length l_i centred on X_i. In that case, instead of the empirical distribution used by the authors to estimate F_0 one might use an absolutely continuous alternative such as the one with density function

$$\hat{f}_0(x)=\begin{cases}\dfrac{1}{l_i}, & x\in L_i, i=1,\ldots,n\\ 0, & \text{otherwise.}\end{cases}$$

In that case

$$\int \log f(x|\theta) d\hat{F}_0(x) = \sum_{i=1}^{n} \frac{1}{l_\cdot} \int_{x \in L_i} \log f(x|\theta) dx$$
$$\simeq \sum_{i=1}^{n} \log f(X_i|\theta) \lambda_i$$
$$= \log \prod_{i=1}^{n} f^{\lambda_i}(X_i|\theta),$$

where $\lambda_i = l_i / l_\cdot$. This would reduce to the usual likelihood when the intervals were of equal length, but that would not always be the case, as when these were interval censored survival times and the intervals could not for practical reasons be made of equal length. The result is the logarithm of a weighted likelihood (e.g., Hu & Zidek, 2002; Wang et al. 2004; Wang & Zidek, 2005). Even if one elects to treat F_0 as non-parametric in a Bayesian context other options are available. In fact, one choice would make F_0 absolutely continuous with respect to Lebesgue measure (Kraft, 1964).

The main point is not whether these alternatives would be better or worse, but rather that the Dirichlet process does not seem canonical *a priori*. As a more concrete example, why not choose the more general Pitman-Yor process? Is it the case that all other choices would not yield the familiar Bayes rule in the parametric families? Or would other choices also yield the familiar Bayes rule but with different matching priors?

The former case would make the choice of Dirichlet process canonical in this context and would be highly interesting. However, it would require a formal treatment. The latter case would shift the burden of selecting a prior to the burden of selecting a predictive non-parametric model. Of course, this burden can be seen as a freedom as well.

The authors bypass absolute continuity issues by arguing that F_0 must be replaced by its expected valued M_0, and later its posterior expected distribution. That step is justified as computing the expected utility and the need to maximize it under the classical normative theory. However, the bypass route ignores an interesting issue that arises, since as first defined in the paper, the optimal parameter value $\theta^* = \theta^*(F_0)$ (our notation) has a prior and then posterior distribution inherited from that of (F_0). In the case of the non-dimensionalized normal distribution for example, $\theta^*(F_0) = \int x dF_0(x)$, a seemingly natural result.

How does that distribution compare to the one derived from loss matching? Or does the phrase "assigning a prior $\pi(\theta)$ which reflects beliefs about θ^*" mean a prior about something different from a prior distribution about the unknown parameter θ^*, for example, about $\theta^*(E[F_0]) \neq E[\theta^*(F_0)]$? The authors rely heavily on the Kullback-Leibler (KL) measure of divergence to characterize loss (or equivalently utility). At the same time, there are a whole class of such measures (the α divergences) that contain KL as a special case. It might be worth exploring the theory in that case. For example, in the simple case where there is a mutual support set equal to $\{z_1, \ldots, z_m\}$,

$$D_\alpha[F_0 || F(\cdot|\theta)] = \frac{1}{\alpha - 1} \log \sum_{i=1}^{m} f^{1-\alpha}(z_i|\theta) f_0^{\alpha}(z_i),$$

where $[f_0(z_1), \ldots, f_0(z_m)]$ has a Dirichlet distribution.

We found the development of the section on loss matching harder to follow. The left hand of the matching equation which is the scoring function loss of $-\log \pi(\theta)$ seems clear enough. The right-hand side seems less intuitive. The first element, which is where the model component of

the prior comes into play, turns out to be Fisher's information $I(\theta)$ based on a heuristic argument that the worth of the model $f(\cdot|\theta)$ indexed by θ increases with the degree of its divergence from its neighbours. The best model would be the one for which this is maximized presumably. But then this utility function has to be turned into a loss function and then put on the log-scale "which is where we are operating" to quote the authors. This meaning and justification for this last step eludes us, especially since Fisher's information is already computed on the log-scale in some sense.

The second element, which is constructed entirely within the model space $\{f(\cdot|\theta)\}$, now assesses models by their predictive performance as measured by a version of KL. Here the true distribution and its prior distribution M_0 component of the full prior specification comes into play and the optimal model would be the one indexed by $\theta(M_0)$ in our notation defined above, where *a priori* $M_0 = E(F_0)$ and *a posteriori* it would be $\theta(M_n)$ with $M_n = E(F_0|X_1, \ldots, X_n)$.

We are not clear about why the two losses should simply accumulate as the authors suggest. Instead we would see the justification as coming from multi-criteria optimization theory. Furthermore, we are unclear as to why the resulting linear combination should necessarily equal the left-hand side. Other criteria could have been used in place of or in addition to those introduced by the authors. Is there a rationality postulate of some kind that forces the equality of the two sides?

Finally, the authors make a good case within their framework as to why $\pi(\theta|x_1, \ldots, x_n)$ should not be found by applying Bayes rule directly to update $\pi(\theta)$. No doubt, this idea will be controversial for it applies to all Bayesian theory that has been developed as part of the program begun a half century ago to find an alternative to the frequency theoretical foundation for statistical inference. It says that since we cannot ever know the true distribution (F_0) of the observable, and that models are always wrong as per George Box's famous maxim, Bayes rule can never be applied directly, instead, updating must always be done indirectly by updating the prior distribution for F_0.

However, as we have argued above, a number of issues need to be resolved before the validity of that conclusion can be accepted and implemented. One important issue is about the discrepancy between the author's findings and those of Akaike indicated above, who finds in favour of direct application of Bayes rule, starting from the same general premises.

Nevertheless, we commend the authors for their thoughtful paper. It is a worthy contribution to the search for a suitable foundation for Bayesian inference.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contrl.*, **AC-19**, 716–723.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika*, **65**, 53–59.

Hu, F. & Zidek, J.V. (2002). The weighted likelihood. *Canad. J. Statist.*, **30**, 347–371.

Jaynes, E.T. (1963). Information theory and statistical mechanics. In *Statistical Physics*, Ed. K.W. Ford, pp. 181–218.

Kraft, C.H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probab.*, **1**, 385–388.

Wang, X., van Eeden, C. & Zidek, J.V. (2004). Asymptotic properties of maximum weighted likelihood estimators. *J. Statist. Plann. Inference*, **119**, 37–54.

Wang, X. & Zidek, J.V. (2005). Derivation of mixture distributions and the weighted likelihood estimator as minimizers of KL-divergence subject constraints. *Ann. Inst. Statistic. Math.*, **57**, 687–701.

[*Received November 2011, accepted November 2011*]