Paradox, simple inconsistency, or serious pathology? Comment on 'Conditioning as disintegration' by Chang & Pollard

M. Stone[1], A. P. Dawid[1], and J. V. Zidek[2]

[1] *Department of Statistical Science, UCL, Gower St, London WC1E 6BT, UK*

[2] *Department of Statistics, University of British Columbia, 6356 Agriculture Rd, Vancouver, BC V6T 1Z2, Canada*

1. Introduction

CHANG & POLLARD (1997) provide a rigorous general formulation of ensembles of conditional probability distributions ("disintegrations") that includes conditioning within general measures. In their Example 12, they apply the formulation to Example 1 of STONE & DAWID (1972) and take the view that it reveals a "flaw" in the marginalization paradox, as expressed in that example. Scepticism and disbelief concerning the general significance of the marginalization paradox as developed by DAWID *et al* (1973) can be found in AKAIKE (1980) and JAYNES (1979) as well as in the discussion of the 1973 paper. The validity of our work was defended in STONE (1982) and DAWID *et al* (1996)—the latter being a response to Chapter 15 of a website publication that has now been revised and posthumously published as JAYNES (2003).

2. Terminology

The term "paradox" was not used in the 1972 paper. Instead, the phenomenon revealed there was described as an un-Bayesian and serious pathology. As a less pejorative description, we might have used "inconsistency", since we now prefer to follow the useful distinction between a paradox (as two aspects of something that appear be irreconcilable but can be seen to be consistent when looked at more closely) and an inconsistency (where such reconciliation is impossible). We regret that this distinction was not drawn and acted on in the major 1973 study, since its use of "paradox" may have encouraged some to ignore the problem as something resolvable by others with the time to think about it.

3. Preparing the ground

Let us agree with Chang & Pollard when they write:

> For a measure λ on the product space $\mathcal{X} \otimes \mathcal{Y}$ it is traditional to use the name $\mathcal{X}$-marginal for the image of λ under the map X that projects onto the $\mathcal{X}$ coordinate space. If λ happens to be a product of probability measures, $P \otimes \nu$, the $\mathcal{X}$-marginal equals P. One can safely refer to both $P\{x \epsilon \mathcal{X} : x \epsilon A\}$ and $(P \otimes \nu)\{(x, y) \epsilon \mathcal{X} \otimes \mathcal{Y} : x \epsilon A\}$ as "the probability that X lies in the

> set A". However, if ν is not a probability measure, the $\mathcal{X}$-marginal of λ does not equal P. At worst, ν might not even be a finite measure, in which case the image measure assigns mass ∞ to every A with $PA > 0$, In this situation there is a real danger in thinking of the $\mathcal{X}$- and $\mathcal{Y}$-coordinates as being independent, or even in thinking of P as the distribution of X.

But the rider that:

> Bayesians with a penchant for improper priors should be particularly aware of this problem.

needs a little elucidation. For example, transgressing notational convention and taking X and Y to be components of the unknown parameter $\Theta = (X, Y)$ for some problem of inference, two cases are to be distinguished assuming that we need to decide, for some other purpose (e.g. inference from some reduction of the data), what to adopt as the prior for X alone. Case 1 is where that prior (the proper P) was the first or given element of λ, which was then augmented more casually with the improper prior ν for Y. Case 2 is where the measure λ was the undifferentiated starting point from which a not-yet-specified prior for X has to be somehow extracted for the alternative purpose. In Case 1, it would be perverse to allow the augmenting ν to affect the status of P as the proper and appropriate marginal prior for X. But Case 2 appears to be as bad as Chang & Pollard's warning suggests. A change of variables from (X, Y) to (X, T) where $T = t(X, Y)$ (of the kind that worried Dempster in the discussion of DAWID *et al* (1973) in which the Jacobian factorizes as $a(X)b(Y)$) will yield a factorization of λ that would typically suggest a quite different marginal prior for X than P.

4. The perceived "flaw" in our Example 1

Example 1 of STONE and DAWID (1972) sets out the problem of a statistician B_1 given two *scientifically certified* exponential random variables X and Y with joint density $f(x, y|\theta, \phi) = \theta\phi^2 \exp\{-\phi(\theta x + y)\}$, where the interest is in inference about $\Theta = E(Y|\Phi)/E(X|\Theta, \Phi)$ (the median of the $\Theta F(2, 2)$ distribution of $Z = Y/X$ that has density $f(z|\theta) = \theta/(\theta + z)^2$). Suppose that prior knowledge of Θ can be expressed with some conviction by the proper density $\pi(\theta)$, whereas hardly anything is known about the magnitude of the common scale parameter $1/\Phi$ of X and Y, as measured on the same uncalibrated instrument. So B_1 (not well-read in the art of simulating ignorance) casually gives Φ a uniform (Lebesgue) measure on R^+.

Chang & Pollard do not question B_1's mechanical derivation of his posterior distribution for Θ: the "(Θ, X, Z)-marginal is sigma-finite" and has an "(X, Z)-disintegration" with a conditional probability density for Θ that is what B_1 got. They do question B_2's posterior density ($\propto \pi(\theta)f(z|\theta)$) simply because they were unable to give it an analogous justification as a "Z-disintegration in the (Θ, Z)-marginal" of B_1's set-up. (The Z-margin is not sigma-finite.) However the inconsistency at the heart of the marginalization "paradox" does not depend on the success or failure of such a justification, because B_2 *starts* his inference engine with the value of Z alone. The notation in our equation (5) did not make this clear enough: it should have been $\pi_{B_2}(\theta|z)$.

Deeper understanding (and a reversal of the pecking order that Chang & Pollard bestowed on B_1 and B_2) comes if we can accept that an unbounded measure to represent knowledge of Φ is at best a convenient approximation to reality. In reality, it would be possible to specify a *lower* bound to the scale factor $1/\Phi$ of the instrument that measured both X and Y—enough to truncate the infinity in the uniform prior for Φ. We should then be able to validate the result of using the uniform prior as a limit, in some sense, of what you can get from proper priors (Stone 1961, 1970, 1982).

This is an old, largely neglected topic on which we will not expand here, except to see what it does for B_1 in Example 1. Instead of truncation, it is mathematically simpler to give Φ an Exp(λ) prior (with proper density $\lambda\exp(-\lambda\phi)$). This gives the posterior

$$\pi_\lambda(\theta|x,y) \propto \pi(\theta)\theta/[\lambda + x(\theta+z)]^3 \tag{1}$$

$$\propto \pi(\theta)\theta/[1 + w(\theta+z)]^3 \tag{2}$$

where $w \equiv x/\lambda$. One sort of limit (satisfying some, e.g. Jeffreys, 1957; Wallace, 1959; Box & Tiao, 1973; Berger, 1980; Leonard & Hsu, 1999) is got if we look at formula (1), fix the data (x,z), and let λ go to zero (to approach uniformity). The limiting distribution is, unsurprisingly, the one that B_1 got.

The other sort of limit comes (if it comes at all) from formula (2) on noting that the influence of λ on how the distribution differs from B_1's is determined by the size of w: only if w is very large can the posterior be approximated by that of B_1. Now the marginal distribution of W is Exp($\lambda\Theta\Phi$) i.e. Exp($\Theta\Psi$) where $\Psi = \lambda\Phi$ is Exp(1). Clearly, under our model, W does *not* in any way go off to ∞ as $\lambda \to 0$. Moreover, because the distribution of W is not degenerate, there is not just one posterior to be considered but a class of very different ones indexed by (w,z). Since the marginal distribution of (W,Z) can be shown to be independent of λ, the probability distribution of the posterior actually realized does not even depend on λ. There is no reasonable sense in which the posterior distributions obtained from any choice of λ can be approximated by that of B_1.

By contrast, this marginal-probability-of-data approach *is* able to justify B_2's posterior if we give Φ a $\Gamma(\lambda,\lambda)$ prior with kernel $\exp\{-\lambda\phi\}\phi^{\lambda-1}$. Then

$$\pi_\lambda(\theta|x,y) \propto \pi(\theta)\theta/[\lambda + x(\theta+z)]^{2+\lambda} \tag{3}$$

$$\propto \pi(\theta)\theta/[1 + w(\theta+z)]^{2+\lambda} \tag{4}$$

With this, the marginal probability that $W \geq w$ is $\int \pi(\theta)(1+\theta w)^{-\lambda}d\theta$, which tends to 1 as $\lambda \to 0$ for any fixed

w. In this case, W *does* go to ∞ in probability as $\lambda \to 0$, and (4) becomes B_2's posterior in the limit. The same posterior is now also given by fixing (x, z) in (3) and taking $\lambda \to 0$—a fortuitous consequence of a combination of underlying group structure and the associated right Haar character of the prior element $d\phi/\phi$ (as explained in STONE & DAWID (1972)).

5. Serious pathology

If there is a flaw in all this, it lies in a willingness (manifested by B_1 in Example 1) to allow an impropriety in a conventionally assigned measure to overwhelm the science embedded in the distribution of Z. Widespread adoption of any argument that lends support to B_1 and finds fault with B_2 would constitute a serious pathology for Bayesian practice. If there are any who doubt this, we would be happy to accept a large number of increasingly large, mutually agreed bets based on B_1's posterior, using Monte Carlo simulation of Θ from an agreed $\pi(\theta)$ and of Z from the corresponding $F(2, 2)$ distribution. That B_1's posterior is not "expectation consistent" (DAWID & STONE, 1973) should be enough to reassure our bank manager.

References

AKAIKE, H. (1980), The interpretation of improper prior distributions as limits of data dependent proper prior distributions, *J.Roy.Statist.Soc.B* **42** 46-52..

BERGER, J. O. (1980), *Statistical Decision Theory and Bayesian Analysis*, p.132 & Sect.3.3.4, Springer-Verlag, New York.

BOX, G. E. P. and TIAO, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Appendix 5.6, Addison-Wesley, Reading Mass.

CHANG, J. T. and D. POLLARD (1997), Conditioning and disintegration, *Statistica Neerlandica* **51** 287-317.

DAWID, A. P. and M. STONE (1973), Expectation consistency and generalized Bayes inference, *Annals of Statistics* 478-485.

DAWID, A. P., M. STONE, and J. V. ZIDEK (1973), Marginalization paradoxes in Bayesian and structural inference (with discussion), *J.Roy.Statist.Soc.B* **35** 189-233.

DAWID, A. P., M. STONE,and J. V. ZIDEK (1996), Critique of E.T.Jaynes's "Paradoxes of Probability Theory", *Research Report 172*, Department of Statistical Science, University College London; also available as http://www.ucl.ac.uk/stats/research/psfiles/172.zip .

JAYNES, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge.

JEFFREYS, SIR HAROLD (1957), *Scientific Inference*, p.68, Cambridge University Press, Cambridge.

LEONARD, T. and HSU, J. S. J. (1999), *Bayesian Methods*, p.135, Cambridge University Press, Cambridge.

STONE, M. (1963), The posterior t distribution, *Ann.Math.Statist.* **34** 568-573.

STONE, M. (1970), A necessary and sufficient condition for convergence in probability to invariant posterior distributions, *Ann.Math.Statist.* **41** 1349-1353.

STONE, M. (1982), Review and analysis of some inconsistencies related to improper priors and finite additivity, *Logic, Methodology, and Philosophy of Science VI: Proceedings of the Sixth International Congress, Hannover 1979* 413-426, North-Holland Publ.Co., Amsterdam.

STONE, M. and A. P. DAWID (1972), Un-Bayesian implications of improper Bayes inference in routine statistical problems, *Biometrika* **59** 369-375.

WALLACE, D. L. (1959), Conditional confidence level properties, *Ann. Math. Statist.* **30** 864-876.