Discussion: Which 'base' distribution?

M.J. Bayarri
Universitat de València

This is a very interesting paper. It addresses a very difficult question for which good answers are sorely needed. We fully agree with the author in his perception of the role of model criticism, and believe that he has provided very useful insights. I, however, disagree with some aspects of the proposed methodology. I focus on the one I perceive as most important, namely, the locating of the realized value of the diagnostic statistic in posterior distributions which are based in the very same data used to compute that realized value. I put my concerns into perspective in the next section. The general developments underlying this discussion are taken from Bayarri and Berger (1999, 2000).

1 Basic issues in model criticism

In the spirit of the paper, assume that we are considering a (null) model $\mathbf{Y}|\theta \sim p(\mathbf{y}|\theta)$, that we observe $\mathbf{Y} = \mathbf{y}_{obs}$ and that we want to investigate the question of whether or not $\mathbf{y}_{obs}$ is compatible with $p(\mathbf{y}|\theta)$. Many common frameworks for model criticism correspond to particular choices of:

(1) A diagnostic statistics $T = T(\mathbf{Y})$ to measure incompatibility of the realized value $t_{obs} = T(\mathbf{y_{obs}})$ with the model.

(2) A 'base', completely specified, distribution of t values, $p(t)$ (under the 'null') in which to 'locate' the observed t_{obs}.

(3) A way to measure conflict between t_{obs} and the base, null distribution $p(t)$. Popular choices include tail areas (as in p-values) and relative hight of the density $p(t)$ at t_{obs}.

The scenario in the paper is more general in that it considers *discrepancy* measures in which T is allowed to also depend on the parameters. Also, conflict is assessed between distributions, and not between a value and a distribution. The scenario above is easier to understand, and will be adopted in this discussion, but virtually identical concerns, questions and potential undesirable behavior arise in the general framework of the paper.

All three choices are very important and determine the power of the selected model criticism procedure. I concentrate only in (2), and discuss the choice of the appropriate 'base' distribution $p(t)$ that we consider optimal for *any* choice of the diagnostic T and *any* way to measure conflict.

2 Distributions for the diagnostic measure

The basic Bayesian tool for model criticism is the prior predictive distribution $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$. Sometimes, however, use of $p(\mathbf{y})$ is unsuitable, as when the prior is improper (or too weak). (Notice that improper priors are most common at the exploratory stages at which model criticism occur.) Instead of abandoning use of $p(\mathbf{y})$ in these circumstances, we propose to use some (suitable) conditional distributions from $p(\mathbf{y})$. That is, choose a *conditioning* statistic $U = U(\mathbf{Y})$, and use as the relevant 'base' distribution for model criticism the *U-conditional predictive distribution*:

$$p(t|u_{obs}) = \int p(t|u_{obs}, \theta)p(\theta|u_{obs})d\theta \ . \tag{2.1}$$

A very popular choice (mentioned in the paper) partitions the data and considers $T = \mathbf{Y_c}$ (the part to 'criticize' the model) and $U = \mathbf{Y_f}$, (the part to 'fit' the model) so that $p(t|u) = p(\mathbf{y_c}|\mathbf{y_f})$. Another popular choice among frequentists takes U to be sufficient for θ (so that $p(t|u,\theta)$ does not depend on θ, and therefore $p(t|u)$ is the same irrespective of the distribution for θ used). Other choices for U are discussed in Bayarri and Berger (1997). Some important remarks are in order:

(1) Because (2.1) is a *valid* probability computation, data does *not* get used twice.
(2) 'Predictive replicate methods' (use of posterior predictive distribution $p(t|\mathbf{y_{obs}})$ as 'base' distribution) can *not* be obtained as a conditional distribution from the prior prior predictive. Indeed, if we take $U = \mathbf{Y}$, $p(t|u_{obs})$ is degenerate.
(3) Once T and U have been chosen, any of the measures of conflict can be used to asses incompatibility of t_{obs} with $p(t|u)$.

Of course, not every choice of $p(t|u)$ is optimal: If we condition too much there might be very little randomness left to assess conflict. If, on the other hand, we condition too little, there might be not enough 'learning' about the parameters. In Bayarri and Berger (1999) we argue that an optimal choice for *any* given T is to use as conditioning statistic, U, the conditional MLE from $p(\mathbf{y}|t_{obs},\theta)$. This however (as the author notices) could give rise to burdensome computations, especially in complicated models, such as the ones typical in HSSS. A very nice approximation (which avoids explicit identification of U) consists in integrating θ out from $p(t|\theta)$ w.r.t. *partial posterior* distribution, defined as:

$$p(\theta|\,\mathbf{y}_{obs}\setminus t_{obs}) \propto p(\mathbf{y}_{obs}|t_{obs},\theta)p(\theta) \propto \frac{p(\mathbf{y}_{obs}|\theta)}{p(t_{obs}|\theta)}p(\theta) \ ,$$

resulting in the *partial posterior predictive distribution*

$$p(\mathbf{y}|\mathbf{y}_{obs}\setminus t_{obs}) = \int p(\mathbf{y}|\theta)p(\theta|\,\mathbf{y}_{obs}\setminus t_{obs})d\theta \ . \tag{2.2}$$

Computations of incompatibility measures with (2.2) are much easier than with (2.1), especially so in cases in which $f(t|\theta)$ is available in close form (in fact, work is in progress with M.E. Castellanos applying (2.2) to criticism of hierarchical models). Moreover, they tend to be very similar (if not identical; see Bayarri and Berger, 1999, 2000). Posterior replicate methods use $p(\mathbf{y}|\mathbf{y}_{obs}) = \int p(\mathbf{y}|\theta)p(\theta|\,\mathbf{y}_{obs})d\theta$, which usually results in even easier computations. Because of this (very compelling) reason, $p(\mathbf{y}|\mathbf{y}_{obs})$ is precisely the 'base' distribution adopted in the paper for model criticism. However the double use of the data involved often results in *severe* conservatism, so severe that clearly unsuitable models are still found compatible with the data, thus rendering the model criticism exercise useless. We present an example in the next section, (taken from Bayarri and Morales, 2001) and many others can be found in previously mentioned references. In fact, when easy of computation is a severe constraint (thus precluding use of (2.2) as base distribution), we recommend use of $p(t|\hat{\theta})$ instead ($\hat{\theta}$ could be the MLE, or the MCMC-able posterior mean), which has proven to produce more satisfactory results than $p(\mathbf{y}|\mathbf{y}_{obs})$ at a sensibly smaller computational burden. (See also the results in Robins, Van der Vaart and Ventura, 2000)

3 An example (Bayarri and Morales, 2001)

As a simple, intuitive example for which the model is clearly contraindicated, consider checking for outliers taking as a 'diagnostic' statistic $T = Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$. The 'null' model is that observations are i.i.d. normal. We simulate 10 observations from a $N(0, \mathbf{1})$ distribution and substitute -8 for the minimum value. The ordered resulting sample is: $\mathbf{-8}$, - 1.27, -1.059, -0 . 986, -0.874, -0.204, 0.315, 0.42, 0.49, 2.457. We use two different measures of conflict (to exemplify that the problem is not with the particular measure of conflict, but rather with the choice of the posterior predictive as 'base' distribution); specifically, for a given 'base' distribution, $p(t)$, we compute: (i) the p-value: $Pr(T \leq t_{obs})$, and (ii) the relative heigh of $p(t)$ at t_{obs}, which we called 'relative predictive surprise', $RPS = \frac{p(t_{obs})}{\sup p(t)}$. We use three different possibilities for $p(t)$: (i) the *partial posterior predictive* distribution (ppp): $p(t|\mathbf{y}_{obs} \setminus t_{obs}) = \int p(t|\theta)p(\theta|\,\mathbf{y}_{obs} \setminus t_{obs})d\theta$, (ii) the *posterior predictive* distribution (post): $p(t|\mathbf{y}_{obs}) = \int p(t|\theta)p(\theta|\,\mathbf{y}_{obs})d\theta$, and (iii) the *plug-in* distribution (plug): $p(t|\hat{\theta})$ (with $\hat{\theta}$ the MLE). The results are shown in Table 3. Because the outlier was 8 standard deviations away from the null mean, it seems rather clear that the only measures of conflict that adequately represent this incompatibility are the corresponding to ppp. Notice how useless the measures based on the posterior predictive are, with a p-value of .133 (and a RPS of .341) even in this very extreme example.

4 Concluding remarks

The approach presented in this paper is a nice generalization of previous approaches, in principle suitably tailored to check models arising from HSSS. However, this nice attempt might fail due to the extreme conservativeness induced by the use of posterior replicate measures. We do agree that computations with such measures can usually be easily incorporated into the regular MCMC

	p-value	RPS
ppp	1.59×10^{-3}	1.87×10^{-3}
post	0.133	0.341
plug-in	0.030	0.128

Table 1 Measures of conflict derived from three different 'base' distributions for the 'diagnostic' statistic

routine, but we feel it very dangerous to use such an extremely conservative approach for model checking, since it gives the users a fause sense of confidence on a maybe disastrous model. The example presented in the paper does find one of the means incompatible with the assumption of exchangeability of all five means, but it should be Noticed that the 'offending' (sample) mean ($\bar{y}_5 = 4.44$) is 17 (sample) standard deviations away from the group formed by the other 4 means(wiht mean 1.5 and S.D. 0.17);

Although the potential for extreme conservatism is the most worrisome, we have also some other minor concerns with the approach taken in this paper. Namely: (i) it is based on the values of densities at particular points, which is known to be not easyly to MCMC-able in general; (ii) it might be difficult to calibrate for all models, all dimensions, etc. (iii) it needs a very carefully specified (and moderately informative) prior distribution. A weak prior will not do; (iv) as a consequence of (iii), the method can not tell apart misspecification of prior opinions from inadequate structure.

In summary, we believe that this paper is a most valuable contribution to a very challenging problem. However, in its present form, it should be used with extreme care: a lack of incompatibility should *not* be taken as an indication that the model is appropriate, but rather that a more powerful checking procedure should be used.

References

Bayarri, M.J., and Berger, J.O. (1997). Measures of Surprise in Bayesian Analysis. *ISDS Discussion Paper* 97-46, Duke University.

Bayarri, M.J., and Berger, J.O. (1999). Quantifying Surprise in the Data and Model Verification. In *Bayesian Statistics 6*, (J.M. Bernardo, J.O. Berger, A.P.Dawid, and A.F.M. Smith, eds.) London: Oxford University Press, pp. 53–82.

Bayarri, M.J. and Berger, J.O. (2000). *P*-values for composite null models. *Journal of the American Statistical Association*, **95**, 1127–1142.

Bayarri, M.J. and Morales, J. (2001). Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference* (to appear).

Robins, J. M., van der Vaart, A., and Ventura, V. (2000). The Asymptotic Distribution of P-values in Composite Null Models. *Journal of the American Statistical Association*, **95**, 1143–1157.