SOME COMMENTS ON MODEL DEVELOPMENT AND POSTERIOR EXISTENCE

We wish to congratulate John Geweke on producing such an interesting and complete paper. We are delighted to see that serious attempts to make Bayesian methods more generally understood and available are being undertaken. In addition, a number of quite challenging issues is addressed here. Whereas we agree with most of what is stated in the paper, it is our (perceived) duty to single out those things that we feel are more contentious. However, we hope that this can lead to a stimulating discussion of general interest and hopefully to our better understanding of these issues.

Model Development

For us, the most interesting novelty of the paper was in Subsection 5.2, where an explicit attempt is made to bridge the gap between classical pure significance testing and formal Bayesian posterior odds analyses. This is a fascinating topic and the literature in this area abounds. Indeed, this is not only a question of deep philosophical interest, but has very tangible consequences for the empirical practice of econometrics and statistics. If we can get away with only partially specifying alternative models and if we don't even need to put a lot of effort in the numerical analysis for the model at hand, before it is "validated" by such an exercise, we can save a lot of time, money and effort.

The analysis proposed here pursues the line of reasoning that was argued by Box (1980) and more recently by Hodges (1998), but the diagnostic is now not based on a classical tail probability, as in Box, who uses prior (within-sample) predictive p-values, but rather upon a Bayes factor with respect to a partially specified model. As such, it tries to bring model development and checking under the Bayesian umbrella, and avoids the mix of paradigms advocated by Box. Clearly, Bayesians can only applaud such an undertaking. Of course, there is a very rich literature on Bayesian diagnostics and model development. Without claiming to be exhaustive, we can mention Zellner (1975) and Chaloner and Brant (1988), who conducted Bayesian residual

analysis, and Rubin (1984), who introduced posterior predictive assessment, extended in Meng (1994) and Gelman, Meng and Stern (1996). Whereas these authors focus on Bayesian diagnostics and do not specify alternative models, the use of posterior odds between fully specified Bayesian models is recommended in *e.g.* the model expansion framework of Draper (1995).

Let us try to understand in some detail what exactly the present paper proposes. Firstly, we are interested in the underlying motivation of partial information Bayes factors. Sofar, we have not been able to find a formal Bayesian justification in terms of a decision theory framework; the latter invariably seems to point towards full posterior odds, based on the predictive for the entire sample (see *e.g.* Subsection 2.3). In the following, we shall use Geweke's notation, and denote by A the complete model under consideration, whereas $\tilde{A}$ is the incomplete alternative model, for which only the prior predictive distribution of a function of the sample $f(Y_T)$ is specified. Partial information Bayes factors, defined as $p[f(Y_T)|A]/p[f(Y_T)|\tilde{A}]$ can be given an interpretation in terms of the usual full-sample Bayes factors in either of the following cases:

a) $f(Y_T)$ contains all the information we retain from the sample. Note that if $f(Y_T)$ is merely a sufficient statistic the above interpretation does not hold. Sufficiency of $f(Y_T)$ means that we can restrict ourselves to the latter for estimating θ, but this does not extend to predictive issues.

b) The incomplete model corresponds to a full-sample predictive given by

$$p[f(Y_T)|\tilde{A}]p[t(Y_T)|f(Y_T), A], \tag{1}$$

where $t(Y_T)$ is such that $[f(Y_T), t(Y_T)]$ is a one-to-one transformation of Y_T. Thus, the implied full-sample predictive would correspond to that of model A for $t(Y_T)$ given $f(Y_T)$, and the alternative model $\tilde{A}$ only differs from A in the marginal predictive of $f(Y_T)$. Whereas case a) is certainly of very limited practical use, case b) might be used to give an implicit justification of partial information Bayes factors in terms of regular Bayes factors.

The second question is then whether the author really has case b) above in mind. This should become clear through the recommended action once the partial information Bayes factor sheds doubt on model A. Do we then look for a fully Bayesian model that generates a predictive that *only* differs for $f(Y_T)$, or do we consider models that would generate the required marginal predictive on $f(Y_T)$, irrespective of their predictive implications for the rest of the sample? Clearly, in the latter case, there is absolutely no guarantee that we will actually improve upon model A in terms of full-sample Bayes factors. Especially since $t(Y_T)$ will typically be of much higher dimension than $f(Y_T)$ the conditional part for $t(Y_T)|f(Y_T)$ could well dominate the Bayes factor. Unfortunately, the author does not give any clear recommendation as to how to proceed once the partial information Bayes factors are computed.

The choice of when to consider a partial information Bayes factor sufficiently in support of further model development is another interesting issue. Clearly, there is an important sense in which a complete model and an incomplete one are not on equal footing. Before careful analysis and further model development we don't know whether we can find any model that leads to the desired predictive $p[f(Y_T)|\tilde{A}]$, and maybe the latter should really be considered a "best-case scenario", providing an upper bound for the evidence against A. On the other extreme, there could be a whole class of convenient models out there that will reproduce that predictive. The prior model probability of $\tilde{A}$ and the cost involved in discarding A could be very different in these cases. What information then does the partial information Bayes factor alone really provide?

Some of the discussion in the previous paragraphs relates to the interpretation that we assign to the incomplete model $\tilde{A}$. Is it meant to be one single model that leads to $p[f(Y_T)|\tilde{A}]$, the set of all the models that can reproduce this predictive, or simply the complement of the currently considered model A? In addition, it is unclear whether we should restrict the predictive implications for the rest of the sample, as indicated above.

Let us now focus on the actual implementation of these ideas in the context of the linear regression model (2.1.7) – (2.1.9) used in the price example analysed by Geweke at the beginning of Subsection 5.3. The statistics considered in the example are based on the least-squares residuals, $u = (I_T - X(X'X)^{-1}X')y$, which take the same value when applied to $\varepsilon = y - X\beta$ instead of just y. Thus, the distribution of u, or any function of it, can never depend on the prior assumed for β. In addition, each of the particular functions $f(u, X)$ considered in the example is scale invariant. As a consequence, the distributions of these statistics will not depend on the precision parameter h either. Thus, in the suggested numerical implementation, we can omit drawing from the prior distribution of β and h. The same results would be obtained by simply assuming that the observables are independent, each following a standard Normal distribution. The resulting partial information Bayes factors are, therefore, entirely unaffected by the prior distribution and, thus, completely ignore an important aspect of the Bayesian model. In fact, rather than reflecting characteristics of the Bayesian model they come closer to providing a diagnostic for the appropriateness of the error distribution for least-squares estimation. Since the prior distribution certainly plays a role in Bayesian inference, we feel that at least some of the choices for $f(Y_T)$ should be able to capture the influence of the prior assumptions.

Pursueing the latter example, some further clarification as to how the incomplete models were chosen would be most helpful. For instance, when John Geweke states that 'one contemplates a predictive distribution for excess kurtosis equivalent to that in the Student-t distribution', it is not clear

to us whether he considers each of the T elements in u to be independently Student-t distributed (with a prior on the degrees of freedom) or whether he assumes such a distribution for the error term ε in the linear regression model and then computes u as the least-squares residuals. The two procedures are clearly not equivalent: whereas the first one leads to a T-dimensional distribution for u, the second one implies a $(T-k)$-dimensional distribution for the latter vector (since it is degenerate in k dimensions). To us, the second method seems more satisfactory, since the incomplete model aims to specify a distribution for $f(u, X)$, where u is, by definition, constrained to a $(T-k)$-dimensional subspace of $\Re^T$. But, in doing this, the incomplete model essentially becomes a complete model (the distribution of the error term is then fully specified). In other words, we would be comparing a Normal versus a Student-t distribution for the error term by means of a partial Bayes factor. This seems to us to somehow defeat the purpose of the exercise, and we would certainly welcome any further light that Geweke could shed on this issue.

In summary, we feel the idea of the partial information Bayes factor is quite interesting and it may well prompt researchers to explore certain directions of model development (by choosing certain $f(Y_T)$'s) and get a quick diagnostic for how much their effort in that direction will potentially reward them. It might well be useful as a quick diagnostic that can serve to reassure the model user that the present model A has no obvious deficiencies, at least when compared with an incomplete model $\tilde{A}$. Of course, the prior predictive p-values of Box (1980) and the posterior predictive p-values of Gelman *et al.* (1996) perform much the same function. There is, however, one important difference: no formal alternative was provided in the predictive p-value approach, whereas Geweke requires the incomplete model, and this can easily give rise to misunderstandings. Certainly in the first example of Subsection 5.3, he comes perilously close to "suggesting" the Student model as $\tilde{A}$ when focusing on excess kurtosis. From merely considering the partial information Bayes factor for kurtosis, one can not conclude that the Student model will fit the entire sample better. As discussed above, a hybrid model as in (1), leading to the Student predictive for kurtosis and the conditional predictive of the Normal model for the rest of the sample would be suggested instead. Adoption of a Student error distribution can only validly be based on a full posterior odds analysis of both models. Given this possibility of misunderstanding and the general lack of clear recommendations, we would hesitate to recommend partial information Bayes factors as a routine procedure at this stage. It is, in our view, not (yet?) as large a step from the analysis in Box as we initially hoped.

We certainly appreciate that this is a very difficult problem indeed, and we applaud the author for tackling it in so gallant a manner. His approach

will certainly add to our understanding of model development and may shed new light on these very intricate issues. We have no novel alternative to offer, but note that, in the absence of clearly defined competing models, predictive calibration against part of the sample, not used for posterior inference, can provide very valuable diagnostic information. Once we are willing to consider other models, our feeling is that there is no real substitute for formal Bayesian posterior odds analysis with fully specified models. Of course, the latter is quite costly, but shortcuts, however informative, inevitably seem to suffer certain drawbacks. Maybe a better understanding of the particular drawbacks involved will allow us to use these methods to our advantage.

Posterior Existence

Whenever one relies on numerical procedures to conduct inference, a careful check of the theoretical feasibility of the analysis becomes a necessity, as is mentioned in the paper on various occasions. In particular, one needs to ensure that the posterior distribution is well-defined and all posterior moments that are computed actually exist. Otherwise, the entire exercise is simply invalid (although Gelfand and Sahu, 1997, investigate proper embedded distributions of smaller dimensions and DeRobertis and Hartigan, 1981, interpret general σ-finite measures to give probability statements in terms of acceptance of bets). Since the issue of posterior existence is crucial for Bayesian inference, we would like to further expand upon Geweke's remarks. The following comments apply to both the cases of proper and improper priors. As mentioned in Fernández and Steel (1998a), we should distinguish two separate issues:

A. The existence of a well-defined conditional distribution of the parameter θ given the observables Y_T. As shown in Florens, Mouchart and Rolin (1990), this holds as long as the predictive distribution is σ-finite, *i.e.* the density $p(Y_T|A)$ in (2.1.4) is finite except possibly in a set of measure zero in the sample space Ψ_T.

B. The fact that the likelihood function for θ corresponding to the observed sample Y_T^0 is integrable (*i.e.* the integral has a positive finite value) with respect to the prior distribution. In that case, $p(\theta|Y_T^0, A)$ defines a p.d.f. for the actually observed sample Y_T^0.

Although **A** and **B** are often thought to be equivalent, it is important to realize that this is not the case. If we have a proper prior $p(\theta|A)$, **A** will always hold, since the joint distribution of (Y_T, θ) is then a probability distribution and, thus, the conditional distribution of θ given Y_T exists. This does not, however, imply that **B** holds for every sample Y_T^0, but it does mean that the set of problematic samples (*i.e.* those samples Y_T^0 for which **B** does not hold) has zero probability under the assumed sampling distribution. In other words, as long as the observations have positive probability under the

sampling model, posterior inference is guaranteed. Note, however, that in the common situation where the sampling distribution is given through a p.d.f. with respect to Lebesgue measure, there can be no guarantee that **B** holds for any given observed sample Y_T^0 that has zero Lebesgue measure. Whereas, under a proper prior, a bounded likelihood ensures that **B** holds, the same need not be true if an improper prior is used. It is worth mentioning, however, that, if the likelihood function is unbounded, there can be improper priors that perform better than proper priors [Fernández and Steel (1998a) present a simple example where an improper prior leads to **B** for any sample Y_T^0, whilst under a proper prior **B** fails to hold for certain samples].

Unfortunately, it turns out that for certain models of empirical relevance the problematic set of measure zero is quite likely to occur in practice. In addition, other undesirable pitfalls can occur: as an example, it could well happen that **B** holds for a given sample, but no longer does when new observations are added (thus destroying the usual updating coherency inherent to the Bayesian paradigm). Fernández and Steel (1998a) analyse these problems for independent sampling from possibly skewed scale mixtures of Normals, and Fernández and Steel (1998b) examine multivariate Student-t regression models. In order to solve these issues, Fernández and Steel (1998a,b) propose the use of set observations, which take into account the precision with which the data were originally recorded. Such set observations have positive sampling probability, allow for posterior inference in most situations of practical interest, and avoid the pitfalls mentioned above.

Most of the paper deals with the case of proper priors. However, many Bayesian studies, both theoretical and applied, make use of improper priors, since the latter often possess interesting "non-informative" or invariance properties. As examples, reference priors (as in Bernardo, 1979) or hierarchical priors are often improper. Improper priors are briefly discussed in Subsection 2.2, in the context of limiting cases of proper priors. To us, this is a somewhat unusual way of introducing improper priors, since limits and integrals (hence, limits in the prior and posterior) can not necessarily be interchanged. In the example used in the paper, a standard Normal linear regression model with known variance and a natural-conjugate Normal prior on β, the regression coefficient vector, the limit considered is obtained by letting the prior variance go to infinity. In the posterior this tends to the result obtained under a flat prior for β on $\Re^k$, which is the limit of the prior kernel, but not of the prior distribution. It can easily be shown that the prior converges in distribution to two point masses, one at $+\infty$ and the other at $-\infty$, assigning probability $1/2$ to each, so even the dominating measure of the prior has changed. Examples of the opposite situation, *i.e.* where the limit of the posterior corresponds to the limit of the prior distribution rather than the limit of the kernel, can also be provided, and we suspect that there

are cases where useful improper prior distributions do not correspond to the limit of any probability distribution at all. Thus, we feel that viewing improper priors solely in the context of limits can lead to misunderstanding.

In the last paragraph of Subsection 2.2, this idea of limits is also applied to the predictive and it is stated that 'an improper prior and a data density do not together provide a joint distribution of parameters and data'. Although to some extent just a matter of terminology, we find the latter statement somewhat misleading: whereas the joint distribution is not a probability distribution, it is still defined, and the only relevant questions are whether **A** or **B** hold. If so, Bayesian inference works just as in the case of proper priors. Viewed from a measure theory perspective, there is essentially no difference between the case of proper and improper priors. Much of the presentation in the paper relies on analytical derivations; we feel that the use of tools from probability or measure theory would allow a more concise and integrated presentation of the Bayesian paradigm. In addition, it would be directly obvious to the reader that Bayesian methods are firmly grounded in probability (or measure) theory and it would, in our personal view, lead to less room for misinterpretation.

We very much agree with John Geweke that questions of posterior existence deserve explicit mention in this paper. A blind application of numerical methods holds great risks of leading us astray. The spectacular advances we have witnessed in numerical techniques (mainly through Markov chain Monte Carlo algorithms) do not free us from careful analytical verification that our numerical results are indeed meaningful. We think it is in the interest of a better understanding and appreciation of Bayesian methods to be explicit and precise in the treatment of these issues. For decades they have been swept under the carpet, certainly in the econometrics profession. We would strongly argue that a proper understanding and application of these matters in particular and probability theory in general is crucial for all those that are involved in the use of Bayesian inference.

Carmen Fernández
Department of Mathematics
University of Bristol
U.K.

Mark F.J. Steel
Department of Economics
University of Edinburgh
U.K.

Additional References

Bernardo, J.M. (1979), Reference posterior distributions for Bayesian inference (with discussion), *Journal of the Royal Statistical Society*, B 41, 113-147.

Chaloner, K. and R. Brant (1988), A Bayesian approach to outlier detection and residual analysis, *Biometrika*, 75, 651-660.

DeRobertis, L. and J.A. Hartigan (1981), Bayesian inference using intervals of measures, *Annals of Statistics*, 9, 235-244.

Draper, D. (1995), Assessment and propagation of model uncertainty (with discussion), *Journal of the Royal Statistical Association*, B, 57, 45-97.

Fernández, C. and M.F.J. Steel (1998a), On the dangers of modelling through continuous distributions: A Bayesian perspective (with discussion), *Bayesian Statistics 6*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford University Press, forthcoming.

Fernández, C. and M.F.J. Steel (1998b), Multivariate Student-t regression models: Pitfalls and inference, *Biometrika*, forthcoming.

Florens, J.P., Mouchart, M. and J.M. Rolin (1990), Invariance arguments in Bayesian statistics, *Economic Decision Making: Games, Econometrics and Optimisation*, eds. J. Gabszewicz, J.F. Richard and L.A. Wolsey, Amsterdam: North-Holland, 387-403.

Gelfand, A.E. and S.K. Sahu (1997), Gibbs sampling, identifiability, and improper priors in generalized linear mixed models, mimeo.

Gelman, A., X.L. Meng and H. Stern (1996), Posterior predictive assessment of model fitness via realized discrepancies (with discussion), *Statistica Sinica*, 6, 733-807.

Hodges, J.S. (1998), Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion), *Journal of the Royal Statistical Society*, B, 60, 497-536.

Meng, X.L. (1994), Posterior predictive p-values, *Annals of Statistics*, 22, 1142-1160.

Rubin, D.B. (1984), Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Annals of Statistics*, 12, 1151-1172.

Zellner, A. (1975), Bayesian analysis of regression error terms, *Journal of the American Statistical Association*, 70, 138-144.