DISCUSSION

P.J. BROWN (*Imperial College, London*):

The paper by Mouchart and Simar gives an elegant exposition of some results consequent on assuming linear Bayes estimates as approximations to Bayes estimates. I have not undertaken a historical search of the relevant literature but I think the results of Hartigan (1969) might be particularly pertinent and the subsequent work of Goldstein (1975a, 1975b, 1976), deserves more direct incorporation.*

In these linear Bayes methods one only needs express the mean and variance of the joint distribution of θ and x. Further features are not required. Of course when the Bayes estimates are actually linear nothing is lost and in this context recently Diaconis and Ylvisaker * have shown how linearity is intimately connected with the exponential family and natural conjugate priors. When one gets away from such situations as is often necessary on various grounds, e.g., the need for fattailed priors to cope with discordant observations (Dawid, 1973 and Hill, 1974) then linear estimates are no longer adequate.

I would be grateful if Professor Mouchart would elaborate on his meaning of 'robustness'. Section 2.2.1 suggests that the less that is assigned the more robust the procedure. Perhaps indeed in the univariate or multivariate situation it is too much to specify all the means and variances. Exchangeability as in Section 2.2 is one way of reducing the problem but should one go further and allow the data to specify the hyperparameters? Efron and Morris (1973) for example introduce Empirical Linear Bayes estimation where, without specifying the distributional forms, linear Bayes coefficients are estimated from the data producing a rather non-linear estimate overall. These then are distributionally relaxed forms of the exciting but non-linear Stein-type estimates.

The interesting paper by Lindley seems potentially rather useful. There are two points that bother me. On Section 1, θ_i (deviation from the maximum likelihood estimate) is said to be $0(n^{-\frac{1}{2}})$. It would be good to have regularity conditions on prior, model and observed data justifying this, especially in view of the nature of the asymptotic expansions and subsequent integrations.

Furthermore, although it is nice notationally to suppress the data, the posterior moments under consideration are functions of both n, the sample size, and the data, so that perhaps more accurate expansions might be available if the data are also

* Incorporated in published version of Mouchart and Simar's paper.

considered. It might even be appropriate in some situations, as for example linear regression, to consider a norm depending on the data rather than just n. At any rate, here, with the usual essential asymmetry of design, n, the sample size, tells us only some of the story.

Overall we should perhaps await numerical comparison with exact results before embarking on these elegant approximations.

M. GOLDSTEIN (*University of Hull*):

Linear Bayes methods are an important recurrent theme in the Bayesian literature (as reflected by the diverse set of references in the paper by Mouchart and Simar). Although I am not quite sure in what sense the authors have simplified previously complicated results, it is useful to have a concise summary of some of the basic work in this area, and I have no particular technical points to make (essentially, I agree with the authors' presentation). Instead, I would like to raise a nontechnical point which puzzles me a little, namely in what, if any, sense can linear Bayes rules be said to be robust? This robustness is stressed at various points by the authors (and by others - I may have done so myself). However, all we are really saying is that the estimator and risk do not depend on many aspects of the prior distribution. But we might, and perhaps should, argue that if different plausible specifications of the full prior distribution give different estimates, then clearly the form of the prior is important, and this aspect of the problem cannot be ignored. Thus, our most conscientious specification of a full prior distribution should give a more meaningful answer than the linear rule, which may not approximate well to any of our plausible range of prior beliefs. I feel that the usefulness of the linear rules is that they say something precise and simple by carefully limiting the aspects of the problem allowed for consideration, but that it would be wrong to attribute any further properties to this approach without careful justification. Do the authors have any comments?

While it is often interesting to perform series expansions of akward integrals, and pick out important terms, to make the claim that the first term omitted is of $O(n^{-2})$ needs careful justification. Thus, if we are evaluating

$$\int_{-\infty}^{\infty} \omega(\theta)e^{L(\theta)}d\theta,$$

clearly this integral is not equal to the integral that $^{-\infty}$ we would obtain by replacing $\omega(\theta)$ and $L(\theta)$ by their respective series expansions. (Indeed, the latter integral may not even converge). For the suggested expansions to work, we must hope that we can find some value $\alpha(n)$ of $O(n^{-1/2})$ for which

$$\int_{|\theta|>\alpha(n)} \omega(\theta)e^{L(\theta)}\, d\theta$$

is of $O(n^{-2})$. Having done this, we may replace $\omega(\theta)$, $L(\theta)$ by their series expansions and retain only the leading terms. This will provide a valid evaluation of the integral between $\pm\,\alpha(n)$ to $O(n^{-2})$. However, even here, as the integrals of the leading terms are not evaluated between $\pm\alpha(n)$ but between $\pm\infty$, we must further check that the difference between the integrals of the leading terms between the different sets of limits is also of $O(n^{-2})$. I suspect that we may be able to do this in many useful cases, but it is not a question of regularity conditions so much as of rate of convergence. As a perhaps slightly unfair question, can we have some guidance as to when these conditions will hold?.

The easy way to explore an approximation is by trying it out for simple problems in which it is straghtforward to evaluate the integral and the approximation. A simple case which, I feel, yields some insight into the procedure is to suppose that we are drawing a sample of size n from a Bernoulli distribution with parameter θ, where the prior distribution for θ is a beta distribution with each parameter equal to a common value γ. As the approximation procedure essentially estimates the "correction" which should be applied, in large samples, to the maximum likelihood estimator $\hat{\theta}$ in order to obtain, approximately, the posterior mean, a natural way to assess the approximation is to consider the ratio

$$r = \frac{\text{actual correction}}{\text{estimated correction}}$$

i.e. the ratio $(\hat{\theta}-E(\theta\,|\,\text{data}))/(\hat{\theta}-\tilde{E}(\theta\,|\,\text{data}))$, where $\tilde{E}$ is the suggested approximation to the posterior mean.

In this case, evaluating the required quantities gives

$$r = \frac{n}{n+2\gamma}$$

(One reason for choosing this example is that r does not depend on the observed number of successes, k, which facilitates a further comparison I shall make below).

Clearly r is of the right order, and as long as n is large compared to γ the approximation will work well. Also the correction is always in the right direction, though it always overestimates the values. However, there is a further, perhaps surprising, interpretation of r for this example, which may illuminate the relationship between the asymptotic approximation and the linear approximations discussed in the paper by Mouchart and Simar. In this problem, the posterior mean is the linear Bayes rule, i.e. the best rule of the form $a\,(k/n) + (1\text{-}a)E\theta$. The value a in this case is precisely the value r given above. Qualitatively, this gives an insight into the range of application of the two approximations. The asymptotic approximation is useful when r is near one, i.e. when a is near one and the linear Bayes rule is near θ. Thus, when n is such that the linear Bayes rule gives negligable weight to the prior mean (i.e. specification of the prior mean conveys very little conformation about the posterior mean), then it is the derivatives around θ which convey useful prior information about the posterior mean. Two further (unfair) questions. Firstly, do these qualitative insights extend to more complicated circunstances, and in particular to multiparameter problems? Secondly, should I be surprised that r is precisely equal to a (i.e. what basic property of the example I chose made it work)?.

J.M. BERNARDO (*Universidad de Valencia*):

Professor Lindley has provided us with asymptotic expansions for often encountered ratios of integrals to order $O(n^{-1})$. Nevertheless, I would like to know more about the question of when n is large enough for the approximation to be used. A formal answer surely depends on the specific problem and on the loss structure attached to the 'distance' between the true value of the ratio of the integrals and its approximation; however, maybe he can give us a feeling of the kind of situations where he expects the approximation to work.

A.P. DAWID (*The City University*):

Professor Lindley's investigation of higher-order approximations to posterior distributions comes at a time of renewed general interest in such approximations for sampling distributions, although I am hard put to recognise the relationship between Lindley's work and the methods of Barndorff-Nielsen and Cox (1979). It seems to me to be more in the spirit of the ideas on second order efficiency considered by Efron (1975). That paper used the idea of *statistical curvature*, a fascinating concept but one which is (as Lindley (1975) himself pointed out in his discussion on Efron) suspect for the Bayesian because of its dependence on the sample space. Nevertheless, I can't help feeling that a parallel, fully Bayesian, theory might be just around the corner, based on a likelihood analogue of curvature, just as the Bayesian first-order theory replaces expected Fisher information by observed information. Such a theory might be valuable for assessing the usefulness of approximations such as those of Mouchart and Simar.

Alternatively, analogues of saddle-point methods might yield accurate non-normal approximations for posterior distributions.

S. FRENCH (*Univesity of Manchester*):

Professor Lindley's paper on approximations to posterior expectations will undoubtedly lead to many fruitful applications. However, before the formulae are used, perhaps one or two cautionary remarks are appropriate.

The approximations required that certain derivatives be calculated and Professor Lindley suggests that the necessity of some rather horrendous differentiation can be avoided by recourse to finite difference approximations. Now, whilst it is generally easier to differentiate than to integrate a function *analytically*, the reverse is true of *numerical* differentiation and integration. Numerical differentiation is a very unstable operation, since it requires many small differences of function evaluations and so rounding error accumulate dramatically. See, e.g. Fröberg (1969), Fox and Mayers (1968) Blum (1972). Since these formulae require the functions to be differentiated numerically three times, these remarks are all the more appropriate.

Therefore, I would suggest that, when Professor Lindley's formulae are used, the functions should be differentiated analytically if at all possible. There are, after all, computer packages that will handle the algebraic operations of differentiation and provide the analytic form of the differential for the vast majority of the functions that arise. If analytic differentiation really is too difficult, then I suggest a visit to one's friendly neighbourhood numerical analyst. We complain enough of non-statisticians doing statistical analyses without consulting us, perhaps we should heed our own advice and consult the experts in numerical analysis.

I.J. GOOD (*Virginia Polytechnic and State University*):

Some of the mathematics in the paper resembles that used in the centroid method of integration of a positive function of several variables. Taylor's theorem in several variables is used and leads to the requirement of calculating the moments of the region of integration. See Good & Gaskins (1969,1971) and Good & Tideman (1978).

In one of Professor Lindley's expansions the term of order $1/n$ was appreciable compared with that of order $1/n^{1/2}$. This suggests that he should take the expansion at least to the next term to check its accuracy in this case.

J. GREN (*Econometric Institute, Warsaw*):

I would like to make a short comment on the paper by Professor Lindley. We know that the problem of multi-dimensional integration is the most difficult problem in Bayesian estimation of econometric models.

Up to now we have two main approaches or directions, to solve this difficult problem.

The first way is just improving the numerical methods for each separate multi-dimensional integral. They seems to be rather unpromising, even for the Cartesian product rule with Newton-Cotes Quadrature at each step.

The second way is to adopt the Monte Carlo method in order to estimate the value of each integral which appears in Bayesian estimation of econometric models. This is much more promising; see Kloek and Van Dick (1978).

Professor Lindley is now proposing a very good and operational approximation for the ratio of multi-dimensional integrals.

Since the ratio of such integrals plays a crucial role in Bayesian estimation technique, Lindley's paper opens a new, third way for obtaining practical results in Bayesian econometrics.

I would like to congratulate Professor Lindley for showing to us this new, very promising method.

A. O'HAGAN (*University of Warwick*):

Professor Lindley's expansions are extremely interesting and promise to become a standard technique, particularly in models with many parameters. For although then the expansions contain a great many terms, the saving over the vast number of function evaluations required for numerical integration will be enormous. The only lingering doubt here is whether the neglected $0(n^{-2})$ terms, whose number will also escalate rapidly, will cease to be negligible.

On a point of methodology it would seem most sensible to expand about the posterior mode $\tilde{\theta}$ than about $\hat{\theta}$. In his univariate example, with the sample from a t distribution, Professor Lindley uses the expansion about $\hat{\theta}$, and with the proper prior he obtains the approximate value of .395 for the mean $\bar{\theta}$. Yet if we expand about $\tilde{\theta}$, using equation (11) rather than (10), we find the new approximation .415. Which is better? Numerical integration confirms the mean to be .415 to three decimal places!

REPLY TO THE DISCUSSION

MOUCHART, M. (*Université Catholique de Louvain*) and SIMAR, L. (*Facultés Universitaires Saint-Louis, Bruxelles*):

Two types of topics seem to emerge from the discussion. The usefulness of Least Squares Approximation in Bayesian Analysis and the claim for robutsness of such procedures.

Let us first mention that the aim of the paper was not to justify the use of L.S. approximation: we only wanted to propose a simple and self-contained exposition of a host of results widespread in the literature. If justification was at stake, two types of arguments could be mentioned. One argument would be to consider the cases where the posterior expectation is (exactly) linear in x. In this connection, as pointed out by P.J. Brown, works like that of Diaconis and Ylvisaker (1979) should be mentioned. Another argument would be to consider whether a given situation is *close* to such a case. As pointed out by A.P. Dawid, the work by Efron (1975) appears to be relevant, in particular, it may help to appreciate the proper role of the coordinates in the choice of parameters and of statistics. In this line of thought, M. Goldstein seems to pay a special attention to the specification of the prior distribution. Although this is surely crucial, we like to insist that the structure of the problem is given by the *joint* distribution of (θ,x). Even if the sampling process is kept fixed, the question of whether the structure of the prior distribution will determine $E\,(\theta\,|x)$ to be more or less linear in x depends on the choice of coordinates.

P.J. Brown raised the question of whether, for example, fat-tailed prior distribution would endanger the use of L.S. approximation. Surely, fat tails in the prior distribution may lead to infinite Bayesian risk (under quadratic loss): in other words the problem may become meaningless. Remember that in a decision context only the product of the prior distribution and the utility function is relevant, thus the prior distribution and the utility function should be specified and discussed jointly. Even if the tails of the prior distribution are rather thick (the variance remaining finite), the linearity of $E\,(\theta\,|x)$ may not be affected; for instance in a multivariate student distribution, the regression functions are still linear; in any cases, $V(\eta)$ will always give an indication on the accuracy of the approximation.

Finally any discussion on *justifying* the use of L.S. approximations should involve the problem of robutsness. This is the second theme of the discussion.

First, a comment by P.J. Brown induces us to clarify a possible misinterpretation of section 2.2.1. Exchangeability was introduced as a generalization of i.i.d. processes. As such it appears as a minimal assumption that allows the reproduction and unification of earlier results; for this reason the hypothesis was decomposed into two steps. Apart from this purely formal aspect, exchangeable but not i.i.d. processes naturally appear e.g. in sampling from finite populations or when integrating out nuisance parameters in i.i.d. processes. In the latter case, indeed, $D\,(x\,|\,\theta_1)$, the data density marginalized on θ_2, a nuisance parameter, represents an exchangeable process and the results of section 2.2 may then be used to evaluate $\hat{E}(\theta_1|x)$ and to obtain in this way a more robust procedure.

Let us now discuss briefly what we mean by the robutsness of L.S.. approximations: this was indeed questioned by both P.J. Brown and M. Goldstein. These approximations act as *smoothing* procedures, i.e. they are less variable from (some) perturbations of a given problem (here, $D(x,\theta)$) than the exact solution. Apart from the search for computational simplification, a possible motivation may be the following: in case of a misspecification error it may be hoped that the approximate solution to a misspecified model might be better than the exact solution of this misspecified model. In any case this approximation is known to be free from systematic error ($E(\eta) = 0$) and some idea of the accuracy may be obtained in a simple way (by computing $V(\eta)$).

Finally we want to thank the discussants: their remarks provided help and opportunity in improving the presentation of our paper.

D.V. LINDLEY (*University College London*):

The most important point made by the discussants concerns the lack of rigour in the derivation of the results and the resulting vagueness about when the approximations are likely to be useful and accurate. These criticisms are correct and their implications most important; but I have to admit that I don't see how to meet them. It is notoriously difficult to assess the accuracy of any expansion without deriving some information about the magnitude of the terms neglected. Even to obtain the term of order n^{-2} would be a formidable undertaking since it would involve the evaluation of R^* (equation (2)) plus other complicated terms. The lesser point of knowing just when the expansion is valid is easier but is beyond my limited mathematical abilities. My feeling at the moment is that understanding will be improved by investigating numerical cases and comparing the exact and approximate results. In this context, I am grateful to O'Hagan for evaluating one integral exactly, with the superb result that it agrees with the approximation to one part in 400. But one swallow does not make a summer and much more investigation is required. We have to be careful too in thinking that a term of order n^{-1} is necessarily less than one of order $n^{-1/2}$. The numerical illustration of the F-distribution in section 2 provides an example to the contrary; and other calculations that I have performed with the Weibull distribution (not reported in the paper) suggest that this can easily happen when skewness is present. The example that most interest me is that of the analysis of variance in Section 3 - and its possible extension to more complicated, higher-dimensional analyses. There it is not quite clear what is the n in the expansion in powers of $n^{-1/2}$, for there are two sample-size parameters; m, the number of groups, and n, the number of observations in each group. Presumably both have to tend to infinity, but does their relative speed of approach matter?

There is one comment on the approximation that can be made with some confidence; the expansion about the mode is typically better than that about the maximum likelihood value. This can be seen clearly in the case discussed by Goldstein. Using the latter he obtains a measure of quality of the approximation equal to $r =$

$n/(n+2\gamma)$. With the modal value, I find r to be $[n + 2(\gamma\text{-}1)]/(n+2\gamma)$. As he points out, r for the likelihood value is near to the desirable value of unity only when γ is small in comparison with n. The modal approximation only requires $n + 2\gamma$, a measure of the total information, likelihood plus prior, to be large. This observation is supported by evaluations of the next non-zero terms in the expansions, which is not too difficult in this case.

Similar remarks apply to the work of Dunsmore: he uses the modal expansion and his results are superior to those using a likelihood expansion. In particular, the latter can easily give rise to negative values when approximating a predictive distribution whereas this only happens for very small samples when using the modal values.

Too much should not be deduced from these calculations since we are here dealing with members of the exponential family, which is unusual in that the derivatives of the log-likelihood above the first are data-free. It is my guess that the results are likely to be most useful in the case where no sufficient statistics of low dimensionality exist and where the sample precision varies from sample to sample. This is why, to enlarge on Dawid's remark, the relationship, or lack of it, between the work of Barndorff-Nielsen and Cox and the results of this paper is of interest; they average over sample values and thereby lose sight of the fact that some samples are more informative than others.

I am grateful to Good for drawing my attention to the centroid method. The main difference between it and the device used in the paper is that the centroid uses a Taylor series expansion of a function, whereas I use one of the logarithm. As a result, where the centroid has moments of inertia, I have moments of distributions. The logarithmic expansion may be preferable here, where it is a sum of n terms, but the two methods are nicely complementary.

I have to confess that a visit to my friendly neighbourhood numerical analyst, as suggested by French, had not occurred to me since I did not see any problems in the evaluation of the differences that could not be solved by intelligent trial and error investigation of the log-posterior in the neighbourhood of the maximum. I did think of analytic differentiation on a computer, but previous experience with this was not encouraging. My personal predilection is for simple numerical analyses using simple computers where I have the feeling, perhaps erroneous, that I know what is going on. Packages and big computers terrify me. They are like some bureaucratic machine where workings and output are unintelligible; like the communication I had from the U.S. Internal Revenue Service which was most unclear as to whether I owed them money or they owed me. Only the subsequent arrival of a cheque clarified the matter. So far as computers are concerned, Schumacher is right; small is beautiful.

Dawid's suggestion of a likelihood analogue of curvature is intriguing. Efron's ideas are useful in discussing the merits of different estimators. But in the Bayesian approach there is only one estimator, the posterior distribution; or, if a decision problem is involved, the unique set of best acts. Hence no optimality considerations arise and thus there appears to be no need for curvature.

Brown is right to raise the question of dependence on the sample. As n increases, additional sample values are introduced, so that any detailed consideration of the limit as $n \to \infty$ must consider how the sample could change. Two possibilities are that we would have asymptotic convergence with probability one for each value of θ: or more weakly, with probability one - this being the overall probability incorporating π (θ).

I do not know the answers to Goldstein's questions. The second involves a very special situation which is perhaps only a curiosity. The first is important because it is in multiparameter problems that the ideas put forward might be most useful.

I am most grateful to all discussants for their sympathetic reception of what is an untidily, incomplete paper.

REFERENCES IN THE DISCUSSION

BLUM, E.K. (1972) *Numerical Analysis and Computation,* Reading, Mass. Addison-Wesley

DAWID, A.P. (1973) Posterior expectations for large observations. *Biometrika,* **60**, 664-666.

EFRON, B. and MORRIS, C. (1973) Stein's estimation rule and its competitors - an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117-30.

EFRON, B. (1975) Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3**, 1189-1217.

FOX, L. and MAYERS, D.F. (1968) *Computing Methods for Scientists and Engineers.* Oxford: University Press.

FRÖBERG, C.E. (1969) *Introduction to Numerical Analysis* (2nd edn) Reading, Mass: Addison-Wesley

GOLDSTEIN, M. (1975a). Approximate Bayes solutions to some non-parametric problems. *Ann. Statist.* **3**, 512-517.

— (1975b). A note on some Bayesian non-parametric problems. *Ann. Statist.* **3**, 736-740.

— (1976). Bayesian analysis of regression problems. *Biometrika* **63**, 51-58.

GOOD, I.J. and GASKINS, R.A.(1969) The centroid method of integration. *Nature* **222**, 697-698

— (1971) The centroid method of numerical integration. *Numerische Mathematik* **16**, 343-359.

GOOD, I.J. and TIDEMAN, T.N. (1978) Integration over a simplex, truncated cubes, and Eulerian numbers. *Numerische Mathematik,* **30**, 355-367

HARTIGAN, J.A. (1969). Linear Bayesian methods. *J. Roy. Statist. Soc. B,* **31**, 446-454.

HILL, B.M. (1974) On coherence, inadmissibility and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics,* (Fienberg, S.E. and Zellner, A. eds.) Amsterdam: North-Holland.

KLOEK, T. and VAN DIJK, H.K. (1978). Bayesian estimates of equation system parameters. An application of Integration by Monte-Carlo. *Econometrica,* **46**, 1-19.

LINDLEY, D.V. (1975) Comments on Efron (1975) *Ann. Statist.* **3**, 1222-1223.