DISCUSSION

JUAN FERRÁNDIZ *(Universitat de València, Spain)*

First of all I would like to thank Professor Casella for this stimulating paper. I have enjoyed reading these many good ideas exposed in so clear a style. I found his main message very important telling us that not only statistical practice can benefit from Markov Chain Monte Carlo (MCMC) methods but that these MCMC methods can still take advantage of well-known statistical ideas.

His second message, related to the Bayesian-frequentist controversy, has been particularly pleasing to me. I strongly agree with Professor Casella that

> "... there are situations and problems in which one or the other approach is better-suited, or even a combination may be best, so a statistician without a command of both approaches may be less than complete."

In fact, as I was reading the paper, I was thinking how his suggestions could apply to a frequentist context: likelihood methods for spatial models arising from random variables associated to geographical sites (see e.g. Ferrándiz *et al.* , 1995).

Gibbs distributions are a natural choice in this context. Among them, the proposed *automodels* in Besag (1974) are particularly appealing because the full conditionals determining joint distributions are well-known members of the exponential family.

The corresponding density of these models can be expressed as

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{t}'\boldsymbol{\theta})h(\boldsymbol{x})}{c(\boldsymbol{\theta})} \tag{1}$$

through a suitable sufficient satistic $\boldsymbol{t}$, where the normalizing constant $c(\boldsymbol{\theta})$ is difficult to compute by standard numerical methods. This fact causes major problems on any inferential procedure based on the likelihood function (including Bayesian posteriors from any prior).

Geyer and Thompson (1992) propose estimating the ratio of constants

$$d(\boldsymbol{\theta}) = \frac{c(\boldsymbol{\theta})}{c(\boldsymbol{\theta}_0)} = E[\exp(\boldsymbol{t}'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) \mid \boldsymbol{\theta}_0]$$

by means of

$$\widehat{d(\boldsymbol{\theta})} = \frac{1}{n}\sum \exp(\boldsymbol{t}_i'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) \tag{2}$$

from a Markov chain simulation $\{\boldsymbol{x}_i : i = 1, \ldots, n\}$ of $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)$. We can then estimate the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ in (1) up to a constant $c(\boldsymbol{\theta}_0)$.

Compatibility of Full Conditionals. Spatial automodels were proposed by Besag (1974) in his pioneering work after he considered the compatibility of full conditionals in order to establish well-defined spatial models. For a finite number of sites and under the positivity condition (the support of the joint distribution equals the product of supports of the full conditionals) we have only to check summability of the joint density. This is not always easy to verify theoretically and it would be very interesting to develop statistical techniques to detect lack of summability directly from the output of the simulation algorithm. A first approach could be to run the algorithm several times from random starting points and check the homogeneity of the produced outputs in the long run. Example 4 in Section 3.2 probably would fail to show any anomalous behavior. I think this is an interesting problem that deserves further research.

Another interesting area of research could be how to relax the positivity condition, which seems quite restrictive in some circumstances like, for instance, when we consider temporal concatenation of spatial distributions in order to build space-time models. It would also be the case, in the Bayesian context, when particular combinations of values of the random variables in the model are impossible.

Rao-Blackwellization. The main difficulty in the likelihood estimation approach for spatial models based on (2) above is the strong variability of $\widehat{d(\boldsymbol{\theta})}$ as $\mid \boldsymbol{\theta} - \boldsymbol{\theta}_0 \mid$ becomes moderately large, producing a useless estimate of the likelihood function outside a small neighborhood of $\boldsymbol{\theta}_0$. The exponential form of the terms in the rhs of (2) make the extreme outliers of the simulated sequence $\{\boldsymbol{t}_i : i = 1, \ldots, n\}$ dominate the sum.

This is a case where it would be worth considering the statistical processing of the output of the simulation algorithm in order to improve our likelihood estimates.

Gibbs sampling is easily implemented in this context because full conditionals $p(x_i \mid \boldsymbol{x}_{-i}, \boldsymbol{\theta})$ are well-known distributions, and no acceptance-rejection mechanism is present. I can not see how the Rao-Blackwellization proposed by Professor Casella in §4 could be applied.

Perhaps, in this case, a robust estimator of the mean could be a good alternative.

Rao-Blackwellization, as proposed in Section 4, seems limited to acceptance-rejection algorithms, where ancillary uniform random variables are used. Gibbs sampling can be stated as a particular case of Metropolis-Hastings algorithm, but with probability one of accepting every move, so that it is not possible to benefit from conditioning on the accepted values in the corresponding accept-reject process. Neither does it seem feasible to apply the ideas proposed in Section 5.2 of Rao-Blackwellising a data augmentation sampling scheme. For this to be done we need a convenient decomposition $(\boldsymbol{t}, \boldsymbol{s})$ of the observed vector $\boldsymbol{x}$ in order to alternate sampling from $p(\boldsymbol{t} \mid \boldsymbol{s}, \boldsymbol{\theta}_0)$ and $p(\boldsymbol{s} \mid \boldsymbol{t}, \boldsymbol{\theta}_0)$. This is not an obvious task.

Nevertheless, I think that the research lines proposed by Professor Casella are very promising. MCMC methods allow the growing complexity of the statistical models considered, and more complex Metropolis-Hastings algorithms are being used. Gibbs sampling has a poor mixing performance in high dimensional problems (as is usually the case of geographical data) and more sophisticated algorithms are being proposed (see e.g. Geyer and Thompson, 1995). The development of statistical treatments of their output has to be welcome as a means to strengthen their utility.

Inference from the Algorithm. On the other hand, the suggestions exposed in Section 5.1 seem worth exploring in the problem at hand. In fact, when we are trying to maximize a log-likelihood function estimate based on (2),

$$\widehat{\ell(\boldsymbol{\theta} \mid \boldsymbol{x})} = \boldsymbol{t}'\boldsymbol{\theta} - \log(\widehat{d(\boldsymbol{\theta})}) + \text{constant} \tag{3}$$

the ratio of constants estimate $\widehat{d(\boldsymbol{\theta})}$ is mostly determined by the extreme outliers of the simulated sequence $\{\boldsymbol{t}_i : i = 1, \ldots, n\}$. Maximization of (3) to get $\tilde{\boldsymbol{\theta}}$, our estimate of the true maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, will be based only on a few outermost observations $\boldsymbol{t}_i$.

Maybe it could be better to partition the whole sequence into small subsamples $\{\{\boldsymbol{t}_i : i = ra + 1, \ldots, (r+1)a\} : r = 0, \ldots, n/a - 1\}$, from which we could get a sequence of log-likelihood estimates $\{\widehat{\ell(\boldsymbol{\theta} \mid \boldsymbol{x})}_r : r = 0, \ldots, n/a - 1\}$. Their maximization will produce a

sample $\{\hat{\boldsymbol{\theta}}_r : r = 0, \ldots, n/a - 1\}$ of estimates of the true maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. The characteristics of this sample could help in monitoring the maximization process. This is a challenging point whose potential benefits deserve further research.

Bayesian readers can translate the problem above to their favorite framework by just adding the required prior $\pi(\boldsymbol{\theta})$ to the likelihood (1) and trying to find the mode of the posterior.

Decision Theory and Algorithms. This is the idea in the paper that I liked the most: to embed MCMC algorithms in appropriate decision problems. There are many decisions to make when running an MCMC procedure (sampling scheme, choice of estimator, stopping rule, etc.). Professor Casella has illustrated the benefit of this approach in some interesting cases. The relevant aspects in practice will come up once we establish the problem in a complete decision framework that takes into account the consequences of our choices. Although it seems in its first steps, I believe in a quick enriching development of this subject whose usefulness is foreseeable.

DANIEL PEÑA *(Universidad Carlos III de Madrid, Spain)*

When I first read this paper I was very disappointed. I found that I was in complete agreement with the main ideas presented on it and therefore my duty as a referee of playing devil's advocate was a very difficult one. Finally I accepted my limitations to be a good discussant of this paper and decided to say what I really believe: This is a wise paper and I am thankful to the editor of *Test* for giving me the opportunity to comment on it.

From my point of view the paper has three main messages. The first one is that we can become better statisticians by adopting a pragmatic approach in which Bayesian and frequentist inference are seen as complementary rather than adversarial. The second one is that there is a risk that today's computer facilities lead us to forget about the internal consistency of the model we are using. This point is very well illustrated by an example in which we may end up estimating, by Gibbs sampling, a non existent posterior distribution. The third message is that we should apply the statistical analysis we preach to the data generated by a computer algorithm and in this way we can not only improve the present algorithm but also create new better ones.

Professor Casella's point of view is that the Bayesian approach is better for the construction of optimal estimators whereas the frequentist one is better for the global evaluation of their properties. I agree on this point. Conditioning on the data has proved to be a very useful method to build estimators but it is not as useful to evaluate their properties which requires integration over the sample space. The same idea has been expressed in a different way by Box (1960) to explain the complementary role of these two statistical methodologies: we need Bayesian inference for estimation and frequentist inference for model checking.

The advantage of Bayesian inference is that it provides a general framework to combine different sources of information in model parameter estimation. Also, as it is well known, any admissible frequentist estimate has a Bayesian interpretation and the Bayesian approach provides straightforward solution in situation in which classical methods are controversial. To quote just but one example, consider the problem of estimating a vector parameter θ by combining information from two normal random variables X and Y where $E(X) = \theta$, $E(Y) = \theta + \xi$, $Var(X) = \sigma^2 I$, and $Var(Y) = \tau^2 I$. Maximum likelihood leads to the simple estimate $\theta = X$, and $\xi = Y - \theta$, in which information about θ coming from Y is not taked into account in the estimation. Assuming prior distributions $\pi(\theta) \sim N(0, v_0 I)$, $\pi(\xi) \sim N(0, \gamma I)$, and letting $v_0 \rightarrow \infty$, it is easy to show that the mean of the posterior distribution is given by

$$E(\theta|X,Y) = X - (\frac{\sigma^2}{\sigma^2 + \tau^2 + \gamma^2})(X - Y)$$

and this estimate minimizes the Bayes risk and is admissible under weak regularity conditions. A related frequentist solution to this problem, in the spirit of James-Stein shrinkage estimator, has been developed by Green and Strawderman (1991). In particular, as they showed in their paper, this estimate can be seen as an empirical Bayes estimate. In general, sensible shrinkage estimators have a straightforward Bayesian justification whereas their derivation in terms of frequentist inference is not so clear. On the other hand, when testing a model without any specific alternative in mind, that is when we look at our model and data and try to see if our hypothesis and the observed data are compatible, we need to have in mind all the samples that might have been observed if the model

was right. The justification of this is better understood in a frequentist point of view. This duality explains why developments in model criticism have mostly been carried out in the frequentist approach and much of the Bayesian literature in the area has just tried to justify frequentist ideas and procedures. For instance, we can find many examples in which Bayesian estimation ideas have lead to better frequentist procedures but there are very few examples of Bayesian diagnostic procedures which have improved the way we do model checking in practice. Some authors have argued that the Bayesian way to deal with this problem is to transform it in a model selection problem which is solved by computing the posterior probability

$$p(M_i|Y) = \frac{p(Y|M_i)p(M_i)}{\sum p(Y|M_i)p(M_i)}$$

where Y is the sample data and $(M_1, M_2, ..., M_k)$ is a set of possible models to be considered. However this formulations has several problems: (i) sometimes we do not have a set of alternative models and we just want to see if the one entertained can be considered a reasonable approximation; (ii) even if we have several models in mind the present application of Bayes theorem requires that we have a partition of the model space, that is the models must be incompatible. In general this is not the case. This is obvious when some models are nested, as when selecting between a linear or a quadratic regression, but in general if we are considering two alternative non-nested models they usually have some degree of overlap. Sometimes we can avoid the overlap by defining all the possible combinations of cases as in selecting the best set of explanatory variables or in outliers problems in which the number of models is 2^n. However, this partitioning of the model space can not be carried out in a clear way in many situations in which we need to choose between several non nested nonlinear models.

In closing my comments on the first message of the paper I would like to stress my full agreement with the final statement of section 2 that both approaches provide to the statistician a better understanding and a more complete approach to statistics. For instance, Samaniago and Renau (1994) showed that the method to be recommended in a particular application depends crucially on the quality of the available prior information. The conclusion of all this is that both approaches

needs to be taught and both should be present in any graduate training in statistics in either the Master or Ph.D. level.

The second important point made in the paper is that the algorithm approach used in a problem has fundamental repercussions on the statistical inference. In the mixed model presented in the paper, assuming some standard non-informative priors for the variances, the posterior distribution does not exist and the inference we obtain by Gibbs sampling does not make any sense. This result stress the need of a careful assessment of the prior distribution in the multiparameter situation mainly in the case in which have mean and variance parameters. Ibrahim and Laud (1991) have showed that if we use Jeffreys's priors under general conditions in generalized linear models the posterior does exit. The paper gives a theorem for the mixed model that is similar in spirit to the one given here and I would ask the author to comment a little bit more on this relationship.

I have found very interesting the application of the Rao-Blackwell theorem to improve the Accept-Reject algorithm. It is a nice example of using the output of a statistical algorithm to improve it, and I would like to add three other examples to the ones presented in the paper.

The first one is using the information provided by Gibbs sampling to improve the convergence of the algorithm when the parameter space is high dimensional and there exists strong correlations among the parameters. This idea has been used by Justel and Peña (1996b) in outlier regression problems with strong masking. These authors showed that Gibbs sampling will fail in this case (Justel and Peña, 1996a) and devise a procedure in which the first runs from the Gibbs sampling are used to learn about the structure of the problem and to modify the starting condition. In this way this modified adaptive Gibbs sampling converges to a solution whereas the standard algorithm does not. The second one is in resampling methods to compute robust estimators. The present algorithms are based on random sampling, and do not take into account the information obtained from previous drawing or from the structure of the problem. For instance, in regression problems we know that points with X variables close to the mean cannot at the same time be outliers and have a small residual. On the other hand we know that high leverage outliers will have a small residual whatever the value of the response variable. If we want to build robust estimates by sampling it seems to be more efficient than random sampling to use stratified sampling where

the allocation takes into account the likelihood that each strata includes unidentified outliers. Peña and Tiao (1992) showed, in a related problem, that if instead of random sampling we use preliminary information to stratify the observations we can obtain a better procedure. Finally, I believe that the use of time series models in the analysis of the output of sequential algorithm can lead to substantial improvement in judging convergence. In particular the use of multiple time series models in the analysis of the output of a parallel algorithm seems to be a promising area of future research.

In summary a have found this paper very stimulating and full of insights. It gives me a great pleasure to congratulate Professor Casella for this outstanding contribution to our journal.

DAVID RIOS INSUA *(Universidad Politécnica de Madrid, Spain)*

Professor Casella makes a very interesting contribution to the study of relations between statistics and algorithms. This topic is extremely vast ranging from Monte Carlo tests and confidence intervals to resampling methods and the probabilistic analysis of algorithms. Casella has concentrated on the hottest topic in the area, that of Markov chain and Monte Carlo methods.

Since their popularisation in Gelfand and Smith (1990), these methods have had a tremendous impact on Bayesian statistics, facilitating analysis of complex models, far more complex than we would have dreamed of a decade ago. Yet, with practice, we are recognising that life is not as simple as promised. Anyone who has done serious work in the area must have faced some of the many potential problems awaiting. As an example, in an earlier version of joint work with Peter Muller on Bayesian analysis of neural network models, we produced a seemingly sensible posterior described by a nice looking histogram. Many readers and listeners of this work were not able to suggest that the reported posterior was not right. We later discovered a bug in our programs, leading to the, what we believe now, right version of the posterior, see Muller and Rios Insua (1996). Incidentally, that was an example in which some of the MCMC folk theorems did not work. For example, blocking of some of the parameters did speed up the algorithms, but the same did not happen for other groups of parameters. A similar phenomenon happened with marginalisation.

Reflecting on our experience and Casella's paper, three main ideas come to mind. The first one is that *there is a clear need in the field to provide guidelines on reporting computational experiments.* This is becoming more important given the increasing impact of simulation methods in Statistics, and the many phantom posteriors that we are discovering. Perhaps, an updated version of Hoaglin and Andrews (1975), not much followed so far, seems in order. These guidelines exist in other fields like mathematical programming, with very healthy effects.

The second one is that Markov chain Monte Carlo seems like a minefield and *we need some kind of roadmap with suggestions of when to use what.* Of course, we still need much more experience with the methods. Casella's paper is a nice step on uncovering the dangers of using improper priors within MCMC, namely that the posterior may be improper and this may be difficult to detect. One way forward, if, for convenience, we insist on adopting improper priors, could be to use sensitivity analysis, as follows. In many cases, there will be a sequence of proper priors converging to the desired improper prior. We could then compare the output produced with those proper priors and the improper prior. Computationally, the approach would not be too onerous, since we could adopt a sampling-resampling perspective, Smith and Gelfand (1992). Conceptually, the approach would provide a much better exploration of the posterior. Theoretically, the approach also entails a number of interesting problems.

As far as the specific example (Figures 1 and 2) in the paper is concerned, one would have expected much more mass near zero. We could wonder whether the sample sizes used are big enough, or whether there might have been problems with the random number generator used, which typically have problems generating numbers very close to 0 or 1.

The third idea is that in spite of Tierney's (1994) review, *the statistical literature has remained relatively ignorant of the operations research and traditional simulation literature*, on issues like initialisation bias, output analysis and variance reduction, see Rios Insua et al (1997). In that direction, Casella's paper is also a fine contribution analyzing a strikingly powerful conditioning technique for variance reduction, based on variants of Rao-Blackwellization. One could wonder how this technique compares with more traditional output analysis or variance reduction methods, specially in the case of dependent data, rather than with independent data as with the Accept-Reject algorithm.

As a final comment, in consonance with Casella's discussion on the interface between classical and Bayesian approaches, and his suggestion of viewing the output from a Monte Carlo algorithm as data, we would be curious to know whether, in his opinion, Bayesian statistics have much of a role in their analyses, given that in this context we are able to gather endless amounts of data.

JOSÉ M. BERNARDO *(Universitat de València, Spain)*

I have very much enjoyed Professor Casella's exposition, and find myself in basic agreement with most of his points. There are, however, some differences of interpretation that I would like to point out:

1. *Proper versus improper priors.* The disturbing fact that people have published Bayesian posteriors which apparently do not exist, because they are based on undetected null Gibbs chains, may tempt some readers to conclude that this is yet another instance of the dangers of using improper priors and that all will be fine if proper priors had been used in the first place. But this is certainly *not* the case.

What probably happens in the examples described is that the Gibbs algorithm in fact is using an "automatic" *proper* approximation to the assumed improper prior, by selecting points in bounded approximations to the unbounded spaces, mirroring the proper approximation to an improper prior which may usually be obtained by truncation. However, if the prior (proper or improper) does not make sense in the problem at hand, the results are not going to be sensible. A prior which leads to an improper posterior will *never* make sense, but a proper approximation to that prior will not make sense either, even if it technically leads to a proper posterior. Generally speaking, one should not blame impropriety for the unsatisfactory results often obtained in multiparameter situations from the use of naïve "default" priors, —marginalization paradoxes (Dawid, Stone and Zidek, 1972), strong inconsistency (Stein, 1959) or the null Gibbs chains discussed here—, for proper approximations to those priors will not work either. What it is necessary is either to specify a true multivariate subjective prior, what is pragmatically often next to impossible, —and for some people it is even undesirable—, or to use a "sensible" default prior which, in particular, must lead to a posterior for the quantity of interest which is dominated by the data.

In the one-way random effects model discussed in Example 4, the use of the "standard" improper power priors on the variances is a well

documented case of careless prior specification; I would really like to see the example reanalyzed with what I would argue to be the appropriate default prior to make inferences about the variances in that problem, namely the *reference* prior

$$\pi(\beta, \sigma^2, \sigma_\epsilon^2) \propto \sigma^{-C_n} \sigma_\epsilon^{-2} \left[(n-1) + \left(\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + n\sigma^2} \right)^2 \right]^{1/2}$$

where $C_n = 1 - \sqrt{n-1}(\sqrt{n} + \sqrt{n-1})^{-3}$ (Berger and Bernardo, 1992), which, naturally, leads to *proper* reference posteriors, for both σ^2 and σ_ϵ^2, for any sample of size $n \geq 2$.

2. *Bayesian evaluation of improved algorithm.* The idea of using statistical techniques for improving the result from MCMC runs by using more sophisticated estimates than the obvious arithmetic average is certainly appealing, and the results on Section 4 provide a *frequentist* argument for its use, by showing a decrease in the mean squared error.

However, as a convinced Bayesian who would use Gibbs to numerically estimate a posterior I cannot analytically obtain, I wonder what the advantages are from a Bayesian viewpoint. Presumably, one would expect to see an appreciable *reduction of the variation* of the estimated posterior when several Gibbs chains are run with the same data and, say, different starting points. It would be nice to see how this works is the simple $Ga(x \mid \alpha, 2\alpha)$ model discussed in Example 5.

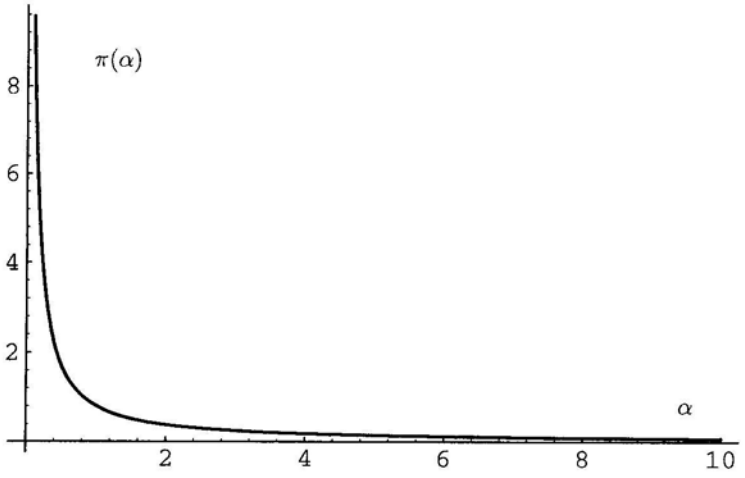

Figure 1. *Reference prior for the parameter of a* $\mathrm{Ga}(x \mid \alpha, 2\alpha)$ *model.*

Of course, the results may depend on the prior used. Since this is a one-parameter regular model, the reference prior is also Jeffreys' prior (Bernardo, 1979), namely

$$\pi(\alpha) \propto \left(- E_{x \mid \alpha}\left[\frac{\partial^2}{(\partial\alpha)^2} \log \mathrm{Ga}(x \mid \alpha, 2\alpha)\right]\right)^{1/2}$$
$$= \left(\psi'(\alpha) - \frac{1}{\alpha}\right)^{1/2} \approx \frac{c}{\alpha},$$

where $\psi'(.)$ is the trigamma, or first derivative of the digamma function, and $c = (\pi^2/6 - 1)^{1/2} \approx 0.65$, shown in Figure 1. It may be seen that, in this case, the refence prior is actually close to the naïve "positive parameter" prior $\pi(\alpha) \propto \alpha^{-1}$.

P. A. GARCÍA-LÓPEZ and A. GONZÁLEZ
(Universidad de Granada, Spain)

We should first like to congratulate Professor Casella for his clear and detailed explanation of all the aspects concerning the interrelationship between statistical theory and computational algorithms, in particular the Gibbs sampler and the accept-reject algorithm. His talk has been highly methodological as far as all aspects of the choice of algorithm and its subsequent effects on the inference are concerned. What we consider to be especially important are the conditions for generating proper posteriors starting from proper conditionals in the Gibbs sampler. Some of the published results on this subject ought to be treated with a degree of caution because the *compatibility* of the proper conditionals (cf. Theorem 2 in Prof. Casella's paper) have not been adequately investigated.

Thus, one question we should like to put to Professor Casella refers directly to a technical aspect of his approach to the application of the Gibbs sampler. There are at least two widely known methods of generating the Gibbs sample, the so-called *single-path* and *multiple-path* methods. Let us suppose that we have a random vector

$$U = (U_1, \ldots, U_k)$$

and that we can simulate the conditional distribution of

$$U_i \mid (U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_k)$$

by using the multiple-path method we draw m independent replicates of the first n cycles of Gibbs samples from the distribution of U, thus obtaining the vector

$$U_n^{(j)} = \left(U_{n1}^{(j)}, \ldots, U_{nk}^{(j)}\right)$$

where (j) denotes the j-th replicate. It is clear that the successive cycles on a particular path, $U_1^{(j)}, U_2^{(j)}, \ldots, U_k^{(j)}$ are not independent but that cycles from different paths, $U_n^{(1)}, U_n^{(2)}, \ldots, U_n^{(m)}$, are indeed independent.

With the single-path method you have only to generate one path long enough to obtain q values for $r + q$, where r is a point at which the Gibbs sampler converges. These q values then provide the basis for our estimation and they all obviously depend upon the starting values.

It has already been demonstrated (cf. Geman and Geman, 1984 and Liu, Wong and Kong, 1992a), that under general conditions, both methods result in convergence, i.e.

$$U_n \xrightarrow{d} U$$

Nevertheless, the dependence between the values generated with the single-path method exerts an influence on the resultant estimators (cf. Gelman and Rubin, 1991).

On the basis of these observations we consider it worth asking Professor Casella the following questions:

1. Are the Gibbs samples in his study based on single or multiple starting values?
2. Has he investigated to see how the choice of cycle values might affect the Gibbs samples thus produced and how this may in turn affect the main result (Theorem 2) with proper posteriors?
3. Do any results exist (similar to those of Theorem 2) for variations of Gibbs sampling as *data augmentation* (cf. Tanner and Wong, 1987) and *substitution sampling* (Gelfand and Smith, 1990)?

To come to another point raised in professor Casella's talk, that of improving the estimators by Rao-Blackwellizing them. It is known that in general the main problem lies in computing the estimators, but there are other, non-parametric methods of improving them, such as the

double-bootstrap. Thus our question is: Have any empirical studies been made to compare the accuracy of the Rao-Blackwellization and double bootstrap methods?

The following contributions were later received in writting.

J. BERGER *(Purdue University and Duke University, USA)*

I congratulate Dr. Casella on a very interesting article. He raises important philosophical and practical questions.

Perhaps the main emphasis of the article is the recommended blending of Bayesian methods (at least regarding MCMC) with frequentist methods. I am certainly also in favor of such, but do have one point of qualification that I think is important. The blendings that Dr. Casella actually uses as examples in the paper primarily involve the use of certain frequentist *tools*, as opposed to the use of frequentist *inferences*. For instance, he demonstrates uses (to a Bayesian) of the law of large numbers and the Rao-Blackwell theorem, two common frequentist tools. Few Bayesians would quarrel with use of such tools (although some might argue that the law of large numbers is as much a Bayesian as a frequentist tool - after all, the first general development of the central limit theorem was by Laplace, and done in an entirely Bayesian way). On the other hand, it is much harder to convince Bayesians that frequentist inferences themselves are of particular use. In the Bayesian's ideal world of the future, numerous frequentist tools will be taught and used, but little in the way of actual current frequentist inference would likely survive. (Many methods that are currently considered to be frequentist, such as maximum likelihood, would still be around, but would be explained as approximations to the Bayesian answers.) Today's frequentists operate in the reverse fashion; they typically admit the considerable value in use of Bayesian tools, but do not find much value in use of Bayesian inferences.

I have a question about Example 4. It has been claimed that, in situations such as this where the impropriety is due to a nonintegrable singularity, the Gibbs sampling output is often reasonable if one does not run the chain for too long. To be more precise, an easy "fix" for such problems is to remove the singularity from the space by, say, introducing the constraint $\sigma^2 > \epsilon$, and the "claim" is that one will often get essentially the same answers from the original Gibbs chain if it is of moderate length

and starts at a reasonable value. This can occur, of course, only if the chain is unlikely to visit too close to the singularity. From the author's experience, is this claim reasonable?

A different issue concerning impropriety, which I have experienced, relates to impropriety due to nonidentifiability. In Andrews, Berger, and Smith (1993) we encountered the fascinating phenomenon that the Gibbs chain for a very high dimensional improper posterior gave convergent estimates for "identifiable" parameters, but not for "nonidentifiable" parameters. This allowed us to determine which parameters were nonidentifiable, and to adjust the model to correct the problem. Taken together with the "claim" in the previous paragraph, this might suggest that impropriety is not necessarily such a concern in hierarchical models; impropriety due to nonidentifiability will be obvious, while that due to singularities is unlikely to affect the answer. Although such a statement verges on sounding ridiculous, we must remember that we are operating in an arena where we will typically never be certain that the Gibbs chain has converged, even if we know that the posterior is proper. Hence all we really need is assurance that, in practice, problems do not seem to arise for the type of problem being considered (e.g., standard normal hierarchical models). While it is fun to speculate about such issues, I must admit that I would not really want to use an improper posterior myself; see also Berger and Strawderman (1996) for additional conditions ensuring proper posteriors in hierarchical models.

Section 4 was quite interesting and had some nice surprises, but I note that it ends up essentially with the "status quo" being supported. The common understanding in use of "accept-reject" and "importance sampling" includes:

(i) Importance sampling gives more accurate estimates for a single h.

(ii) If one wants to simultaneously compute expectations for many h, but the same Y, accept-reject will often be computationally faster, especially if the acceptance rate is low (since then t will be considerably smaller than n).

iii) Rescaling by a correlated estimate of one is an important variance reduction technique.

Use of versions of Rao-Blackwellization does not really appear to add much here. Later examples in the paper do, however, show considerable gain in use of Rao-Blackwellization.

In Section 5.3, I am curious as to whether use of the optimal random scan based on the minimax criterion is actually superior to the optimal scan based on convergence rate (for other than the least favorable function h, of course).

A. P. DAWID *(University College London, UK)*

The general idea of "Rao-Blackwellisation", as a way of improving an inference by eliminating unwanted stochastic variation, is an important and powerful one, as this paper reconfirms. I am surprised it is not used more widely, particularly in its simpler variants. For example, why does any one still do accept-reject sampling (Section 4.1) for Monte Carlo estimation of f based on a sample from g? If the improvement δ_{RB} of δ_{AR} seems over-complex, a simpler approach is just to replace $I(U_i \leq w_i)$ by $E\{I(U_i \leq w_i)\} \equiv w_i$, leading back to the very simple importance sampling estimate δ_I. This is the exact Rao-Blackwell improvement on δ_{AR} when the number N of Y_i's generated from g (but possibly rejected) is fixed, so that the number of retained terms in the accept-reject formula is random. I am not sure of the practical value of Casella's more intricate analysis, which takes into account the randomness in N; and its dependence on the stopping rule offends against some of my deep intuitive feelings about inference. Does this extra complexity have a real pay-off?

A good way of thinking about importance sampling is as follows. We want to approximate the distribution with density f. To do so, we generate points (y_i) from another density g, and to each y_i we attach weight $w_i = f(y_i)/g(y_i)$. We end up with a discrete measure μ_N, having mass w_i at y_i $(i = 1, \ldots, N)$. Normalizing this (by N for unbiasedness, or, better, by $\sum_1^N w_i$ to ensure total mass 1 and thereby improve overall accuracy) to P_N, we get $P_N \to P$, the desired distribution with density f. The expectation of any function under P_N then provides the importance sampling estimate of its expectation under P. From this viewpoint, accept-reject operates by forming an approximating distribution to P by thinning out the (y_i), only retaining y_i with probability proportional to w_i; and attaching equal weights to the retained points. Its inefficiency is self-evident, and that it should have been proposed at all may be

attributed to a subconscious feeling that a discrete distribution must have equal weight on every point—a position that does not stand up to a moment's scrutiny.

Metropolis-Hastings simulation has similar features to accept-reject. Consider a M-H chain with proposal density $q(y' \mid y)$ and acceptance probability $\alpha(y, y')$, satisfying detailed balance for a target distribution P having density f:

$$f(y)q(y' \mid y)\alpha(y, y') \equiv f(y')q(y \mid y')\alpha(y', y).$$

Let $\beta(y) := \int \alpha(y, y')q(y' \mid y)dy'$ be the overall probability of accepting a proposal to move from y. Suppose that we continue until a fixed number N of proposals have been accepted. Ignoring burn-in, thinning, *etc.*, estimation of $\mu := E_P\{h(Y)\}$ by its corresponding chain average is equivalent to estimating P by the (normalised) discrete measure on the successive accepted proposals $x_1, \ldots, x_N$, with x_i being assigned weight W_i, the number of trials starting from x_i before the next proposal is accepted. But W_i is random, with a geometric distribution (conditioned on past x's) having mean $w_i := \beta(x_i)^{-1}$. "Rao-Blackwellisation" thus suggests it would be better to replace the observed number W_i of repetitions of x_i by the new weight w_i (assuming this can be calculated). If we can actually simulate directly from the embedded Markov chain of accepted proposals $x_1, x_2, \ldots, x_N$, with transition density $\gamma(x' \mid x) : = q(x' \mid x)\alpha(x, x')/\beta(x)$, a much more efficient procedure is obtained. If not, and we still have to generate and reject proposals, it should still be more efficient; and it seems likely that still further advantage could be taken of the rejected values, parallel to suggestions of Casella and Robert (1996b).

THOMAS J. DICICCIO and MARTIN T. WELLS
(Cornell University USA)

It is a pleasure to participate in this discussion of Professor Casella's paper on the interplay between Markov Chain Monte Carlo (MCMC) algorithms and statistical inference. The underlying theme of this paper is statistical inference for parameters based on MCMC output. This discussion begins with a few specific questions and then focusses on some relationships between the Casella's minimax decision theory approach of Section 5.3 and the literature on rates of convergence of MCMC methods via the second dominant eigenvalue.

Professor Casella begins with a very welcome call to use Bayesian and frequentist approaches in a complementary way; in particular, his suggestion of using frequentist performance to distinguish between and improve upon estimators that arise from Bayesian considerations is most reasonable. In the context of the popular general linear mixed model, Professor Casella vividly demonstrates some seemingly catastrophic pitfalls that choosing a prior distribution can present. Theorem 1 identifies priors for this model that are appropriate from a Bayesian perspective. A natural question is whether any of these prior distributions produce inferences that are correct or nearly correct from a frequentist perspective. In particular, is there any compelling inferential rationale for choosing $a = b = 1$ in Example 4?

Figures 1 and 2 are certainly startling and distressing from a Bayesian perspective. However, Professor Casella appears to have a firm understanding of their behavior from the underlying "null Gibbs chains." Is it possible that, despite the Bayesian catastrophe, the algorithm could be used to produce reasonable frequentist inferences?

The Rao-Blackwellization and related methods described in Section 4 are ingenious and potentially very useful. It is not unreasonable to consider them from the viewpoint of frequentist inference, given the current interest in noninformative priors and probability matching. Typically, the upper $1-\alpha$ quantile of the marginal posterior density for a scalar parameter of interest is an approximate upper $1-\alpha$ confidence limit having coverage error of order $O(n^{-1/2})$. If a Welch-Peers noninformative prior is used, this error might be reduced to $O(n^{-1})$. If frequentist inference is the ultimate goal, given that the inferences obtained from the exact posterior distribution are at best rather approximate, is there any benefit necessarily to using Rao-Blackwellization? What is the interpretation of Tables 1 and 2 in connection with noniformative priors?

To view Professor Casella's minimax decision theory approach in connection with rates of convergence of MCMC methods and second dominant eigenvalues, some background results and notation is necessary. Let $\{X_j\}$ be a discrete-time homogenous Markov chain on $\mathcal{X}$, with transition probability matrix $P = \{p(x,y) : x,y \in \mathcal{X}\}$, where $P(x,y) = P\{X_j = y \mid X_{j-1} = x\}$. Define the $k-$step transition probabilities by $P^k = \{p(k,x,y) : x,y \in \mathcal{X}\}$. The stationary measure $\pi(x)$ on $\mathcal{X}$ of course satisfies $\pi P = \pi$, that is, $\sum_x \pi(x)p(x,y) = \pi(y)\ \forall y \in \mathcal{X}$. Let $\ell_2(\pi)$ be the Hilbert space of real-valued functions

on $\mathcal{X}$ with inner product $< f \mid g >= \sum_x \pi(x)f(x)g(x)$. The equilibrium expectation of f under π is then $< f >\equiv< f \mid 1 >= E_\pi(f)$, and we can think of $(P^k f)(x)$ and $(\Pi f)(x)$ as operators on $\ell_2(\pi)$ given by $P^k(f)(x) = \sum_y p(k,x,y)f(y)$ and $(\Pi f)(x) = \sum_y \pi(y)f(y)$. The matrix Π has rows equal to π and is an orthogonal projector on $\ell_2(\pi)$ with range the constant functions. The autocovariance function of $\sum_i f(X_i)$ is

$$C_f(\mid i-j \mid) = E_\pi\{[f(X_i) - E_\pi f(X_i)][f(X_j) - E_\pi f(X_j)]\},$$

which also equals $< f \mid (P^{|i-j|} - \Pi)f >=< f \mid (P-\Pi)^{|i-j|}f >=< f \mid (I-\Pi)P^{|j-j|}(I-\Pi)f >$. The autocorrelation function is $\rho_f(|t|) = \frac{C_f(|t|)}{C_f(0)}$.

In Section 5.3 Professor Casella discusses the minimax properties of the Monte Carlo average estimate of the parameter $\mu = E_\pi h(X)$. The limiting risk function $R^{(n)}(h)$ in (23), can be developed further by using the results of Peskun (1973). The fundamental matrix of Markov chains (Kemeny and Snell, 1983) $Z = (I-(P-\Pi))^{-1} = I + \sum_{k=1}^{\infty}(P^k - \Pi)$ arises naturally in this limiting expression. It can be shown that

$$\lim_{n\to\infty} R^{(n)}(h) =< h \mid Qh >,$$

where $Q = 2Z - I - \Pi = (I+P)(I-P)^{-1}(I-\Pi)$. Moreover, by using the series representation of Z and the definition of the autocovariation function, it can be shown that $\lim_{n\to\infty} R^{(n)}(h) = \sum_{k=0}^{\infty} C_h(k)$.

In this case of where $\{X_i\}$ are independent, $\lim_{n\to\infty} R^{(n)}(h) = < h \mid (I-\Pi)h >= C_h(0)$. The ratio

$$\tau_h = \frac{1}{2}\frac{< h \mid Qh >}{< h \mid (I-\Pi)h >}$$

is known as the integrated relaxation time, see Sokal (1989) and Gidas (1995). There are $n/2\tau_h$ effectively independent samples in a run of length n. Note that $\tau_h = \frac{1}{2}\sum_i \rho_h(i)$.

Professor Casella asserts that the risk function contains more information than is contained in the rate of convergence. This assertion can be seen using the ideas above. In the case where P is self-adjoint on

on $\mathcal{X}$ with inner product $< f \mid g >= \sum_x \pi(x)f(x)g(x)$. The equilibrium expectation of f under π is then $< f >\equiv< f \mid 1 >= E_\pi(f)$, and we can think of $(P^k f)(x)$ and $(\Pi f)(x)$ as operators on $\ell_2(\pi)$ given by $P^k(f)(x) = \sum_y p(k,x,y)f(y)$ and $(\Pi f)(x) = \sum_y \pi(y)f(y)$. The matrix Π has rows equal to π and is an orthogonal projector on $\ell_2(\pi)$ with range the constant functions. The autocovariance function of $\sum_i f(X_i)$ is

$$C_f(\mid i-j \mid) = E_\pi\{[f(X_i) - E_\pi f(X_i)][f(X_j) - E_\pi f(X_j)]\},$$

which also equals $< f \mid (P^{|i-j|} - \Pi)f >=< f \mid (P-\Pi)^{|i-j|}f >=< f \mid (I-\Pi)P^{|j-j|}(I-\Pi)f >$. The autocorrelation function is $\rho_f(|t|) = \frac{C_f(|t|)}{C_f(0)}$.

In Section 5.3 Professor Casella discusses the minimax properties of the Monte Carlo average estimate of the parameter $\mu = E_\pi h(X)$. The limiting risk function $R^{(n)}(h)$ in (23), can be developed further by using the results of Peskun (1973). The fundamental matrix of Markov chains (Kemeny and Snell, 1983) $Z = (I-(P-\Pi))^{-1} = I + \sum_{k=1}^{\infty}(P^k - \Pi)$ arises naturally in this limiting expression. It can be shown that

$$\lim_{n\to\infty} R^{(n)}(h) =< h \mid Qh >,$$

where $Q = 2Z - I - \Pi = (I+P)(I-P)^{-1}(I-\Pi)$. Moreover, by using the series representation of Z and the definition of the autocovariation function, it can be shown that $\lim_{n\to\infty} R^{(n)}(h) = \sum_{k=0}^{\infty} C_h(k)$.

In this case of where $\{X_i\}$ are independent, $\lim_{n\to\infty} R^{(n)}(h) = < h \mid (I-\Pi)h >= C_h(0)$. The ratio

$$\tau_h = \frac{1}{2}\frac{< h \mid Qh >}{< h \mid (I-\Pi)h >}$$

is known as the integrated relaxation time, see Sokal (1989) and Gidas (1995). There are $n/2\tau_h$ effectively independent samples in a run of length n. Note that $\tau_h = \frac{1}{2}\sum_i \rho_h(i)$.

Professor Casella asserts that the risk function contains more information than is contained in the rate of convergence. This assertion can be seen using the ideas above. In the case where P is self-adjoint on

$\ell_2(\pi)$, that is $< g \mid Ph >=< Pg \mid h >$, one can relate the rate of convergence of the chain to the limiting risk. Let the ordered eigenvalues of P are $1 = \beta_0 > \beta_1 \geq \beta_2 \geq \cdots \geq \underline{\beta} \geq -1$, where $\underline{\beta}$ equal the smallest eigenvalue. Much work (see Diaconis and Stroock, 1991) has focused on methods for bounding $\beta_1, \underline{\beta}$, and $\beta_* = \max(\beta_1, \mid \underline{\beta} \mid)$ that give rise to bounds on the rate of convergence of the chain to its stationary distribution. As pointed out in Diaconis and Stroock (1991), there are advantages to studying $I - P$ instead of P. The spectrum of $I - P$ consists of numbers $\lambda_i = 1 - \beta_i$. Using the minimax representation of eigenvalues

$$\lambda_1 = \inf_h \frac{< h \mid (I - P)h >}{< h \mid (I - \Pi)h >} = \inf_h [1 - \rho_h(1)] \quad .$$

where the infimum is over all nonconstant functions $h \in \ell_2(\pi)$, this ratio is called the Rayleigh quotient and its numerator may be represented as

$$\frac{1}{2} \sum_{i,j} \pi(i) p(i,j) \mid f(i) - f(j) \mid^2 .$$

The rate of convergence of the chain is determined by λ_1 and hence by the infimum of $\rho_h(1)$ over $h \in \ell_2(\pi)$. Therefore, as limiting risk is essentially a series in $\rho_h(k)$ and β_1 is related to $\rho_h(1)$, the limiting risk contains more information.

Using the ideas above we can study a special case of the random scan. Suppose the transition matrix is a mixture of two transition matrices, that is, $P_\lambda = (1 - \lambda)P_1 + \lambda P_2$. First, it is easy to see that $\lambda_1(P_\lambda)$ is a concave function of λ using the minimax representation. As for $\tau_h(P_\lambda)$, a bit more work is needed. On the orthogonal complement of the constant functions, we have that $Q = 2(1 - P)^{-1} - I$. Using the result of Caracciolo *et al.* (1990) that

$$< f \mid (A^{-1} + B^{-1})^{-1} f >\leq [< f \mid Af >^{-1} + < f \mid Bf >]^{-1}$$

for A and B positive definite self-adjoint matrices, with $A = (1 - \lambda)^{-1}(I - P_1)^{-1}$ and $B = \lambda^{-1}(I - P_1)^{-1}$ it follows that

$$\begin{aligned}(1 - \lambda) < h \mid (I - P_1)^{-1} h > + \lambda < h \mid (I - P_2)^{-1} h > \\ \leq < h \mid [(1 - \lambda)(1 - P_1) + \lambda(I - P_2)]^{-1} h >^{-1},\end{aligned}$$

and

$$(\tau_h(P_\lambda) + 1/2)^{-1} \geq (1-\lambda)[\tau_h(P_1)+1]^{-1} + \lambda[\tau_h(P_2)+1]^{-1}.$$

Hence both $[\tau_h(P_\lambda)+1]^{-1}$ and $\lambda_2(P_2)$ are concave functions of λ. A consequence of this convexity is that

$$\lambda_2(P_\lambda) \geq \min(\lambda, 1-\lambda) \sup_{0<\lambda<1} \lambda_2(P_\lambda)$$

and

$$[\tau_h(P_\lambda) + 1/2]^{-1} \geq \min(\lambda, 1-\lambda) \sup_{0<\lambda<1} [\tau_h(P_\lambda) + 1/2]^{-1}.$$

Hence the randomized approach with $\lambda = \frac{1}{2}$ is never more than a factor of 2 from the best value of λ.

PAUL GUSTAFSON *(University of British Columbia, USA)* and
LARRY WASSERMAN *(Carnegie Mellon University, USA)*

George Casella has presented us with an interesting perspective on the relationship between computing and statistical theory. He makes it clear that the two are inexorably intertwined. Each area enriches and informs the other. He has also emphasized that there is an inevitable mixture of Bayesian and frequentist ideas when one considers statistical computing algorithms and their relationships with inference.

We agree that both Bayesian and frequentist methods are necessary and that statistics is at its best when the two are in happy coexistence. Of course there are many who do not agree on this point and we hope that George's article will help convince the doubters (Bayesian or frequentist) of the need for both.

As should be clear by now, we have little disagreement with anything in this article. We do wish to raise a few points.

1. Averaging Conditional Densities Can Fail. The paper discusses several aspects of the "Rao-Blackwellization" of estimators applied to Monte Carlo output. The author also mentions the "usual average of conditional densities" estimator of a marginal density, which is in the same spirit as Rao-Blackwellized estimators of expectations. For brevity we will refer to this estimator as the ACD (Average of Conditional Densities) estimator. Conventional wisdom dictates that the ACD estimator

of a posterior marginal density is the preferred estimator in any context where it can readily be calculated. We would like to point out a curious and undesirable feature of the ACD estimator in certain hierarchical model settings.

We look at an artificially simple hierarchical model in order to illustrate this feature clearly. Specifically, consider a simplified version of Example 4, where β and σ_ϵ^2 are known, and the prior on σ^2 is locally uniform ($a = -1$). Further, assume that $n_i = 1$ for $i = 1, \ldots, k$, so that we can write Y_i unambiguously. It is simple to verify that the joint posterior distribution on μ and σ^2 is proper. In what follows below, a density for σ^2 evaluated at zero will be defined as the obvious limit.

If the goal is estimation of the marginal posterior density of σ^2, the ACD estimator is

$$p_{ACD}(\cdot) \equiv \hat{\pi}_{\sigma^2|y}(\cdot|y) = \frac{1}{m}\sum_{i=1}^{m} \pi_{\sigma^2|\mu,y}(\cdot|\mu^{(i)}, y), \tag{1}$$

where $\{\mu^{(i)}\}_{i=1}^m$ are the μ vectors sampled by the Monte Carlo scheme. The conditional posterior distribution of $\sigma^2|\mu, y$, which appears on the right-hand side of (1), is inverse gamma, with shape $(k/2) - 1$ and scale $(1/2)\sum_{i=1}^k \mu_i^2$. On the other hand, the true marginal posterior distribution of $\sigma^2|y$ is identical to the conditional distribution of $(T - \sigma_\epsilon^2)|T > \sigma_\epsilon^2$, where T has an inverse gamma distribution with shape $(k/2) - 1$ and scale $(1/2)\sum_{i=1}^k (y_i - \beta)^2$. Thus the true posterior marginal density for σ^2 is finite and positive at $\sigma^2 = 0$. But since the inverse gamma density is always zero at $\sigma^2 = 0$, the ACD density estimate is always zero at $\sigma^2 = 0$, no matter how large a Monte Carlo sample is drawn. In other words, $\pi_{\sigma^2|y}(0|y) > 0$ yet $p_{ACD}(0) = 0$. Thus the ACD estimator is inconsistent at $\sigma^2 = 0$. It might be tempting to dismiss this concern, since it is only an issue at the boundary of the parameter space. But in fact an ACD estimate is going to be misleading about the shape of the posterior marginal density near zero. This is especially true for data sets with $(1/k)\sum_{i=1}^k (y_i - \beta)^2 \leq \sigma_\epsilon^2$. In such cases, the true posterior marginal density for σ^2 takes on its maximum value at zero and is monotone decreasing, which can be interpreted as evidence in favor of $\sigma^2 = 0$. But for any Monte Carlo sample the ACD density estimate will be zero at $\sigma^2 = 0$ and will be increasing on at

least some small interval extending right from zero. This suggests that $\sigma^2 > 0$. Thus the ACD estimator has great potential to be misleading about the posterior evidence concerning small values of σ^2.

This aberrant behavior has been illustrated in a very simple model where the posterior marginal distribution of the variance component can be obtained analytically. The behavior seems to occur quite generally, however, whenever a prior density which is positive at zero is specified for a variance component. The use of such priors seems quite appropriate in many contexts, even though inverse gamma priors, which vanish at zero, are much more commonly specified for variance components. The data cannot rule out the absence of a random effect ($\sigma^2 = 0$), so it seems overly confident to use a prior which vanishes as σ^2 goes to zero. In fact, one might argue that monotone decreasing prior densities should be specified, in order to favor parsimonious models. The Jeffreys prior for the simple model discussed above has a monotone decreasing density which is finite and positive at zero. One disadvantage of *not* using an inverse gamma prior is that the "conditional conjugacy" which drives the Gibbs sampler will be lost. The ACD approach can be extended to deal with this, however, based on work of Chen (1994). But Chen's density estimator will still have aberrant behavior near zero.

In one sense it is not surprising that the ACD estimator does not work well for variance component marginals with prior densities which are positive at zero. In such problems, the Bayes factor for testing the absence of random effects can be expressed as the Savage-Dickey density ratio, which is the ratio of posterior to prior marginal densities for the variance component, both evaluated at zero. For details see Verdinelli and Wasserman (1995). If the ACD estimator worked well for estimating the posterior marginal density at zero, then we would have an easy and reliable way to estimate the Bayes factor. But invariably Bayes factors are harder to compute than other posterior quantities. In this regard, we are not surprised that there is no free lunch via the ACD estimator.

2. *Priors for Hierarchical Models.* As discussed in the paper, choosing priors for hierarchical models is delicate. The dangers of improper posteriors are real and insidious. The theorems reviewed in the paper should prove valuable for guiding statistical practice. However, it seems that many statisticians try to deal with this problem by replacing improper priors with vague proper priors. This merely approximates an ill-defined

posterior with a nearly ill-defined posterior. We would like to mention another solution to the problem.

One output of an inference from a hierarchical model is shrunken estimates. In some cases, conditionally on the hyperparameters, the shrunken estimates lie between the prior mean and the m.l.e's from a non-hierarchical model, i.e. $\hat{\theta}_{\text{Shrunk}} = \alpha\theta_0 + (1-\alpha)\hat{\theta}$, say. It seems reasonable to place a uniform prior on the degree of shrinkage α. This implies a (proper) prior on the hyperparameters. This idea has been used by Strawderman (1971), Christiansen and Morris (1994), Daniels and Gatsonis (1996) and others. It is similar to a prior suggested by DuMouchel (1994). The full generality of the idea is explored in Daniels (1996). This prior seems to be a general way for providing proper reference priors for hierarchical models. Yet another alternative is to place a proper prior (such as half normal or half Cauchy) on the distance from the "null" sampling model in which the random effect is 0. Jeffreys pointed out that such strategies often lead to useful, proper reference priors.

As a general remark we would add that any time improper priors lead to trouble, we should not use vague proper priors. To do so is simply to approximate an ill defined solution. Instead, proper reference priors are called for. Similar problems occur in using Bayes factors to compare models. It is well known that improper priors lead to ill-defined Bayes factors. As Jeffreys made clear, the solution is not to use vague proper priors but rather, to use proper reference priors.

EDWARD I. GEORGE *(University of Texas at Austin, USA)*

Let me begin by congratulating Casella for a masterful paper which synthesizes and interweaves so many different ideas and points of view. There is much to comment on, as Casella seems to open up a whole new vista of ideas with each new section. However, for the sake of focus (and space), I would like to confine my comments to Section 3.2 which is concerned with the properties of Gibbs Markov chains when the Gibbs conditionals do not correspond to a proper posterior.

The key result of Section 3.2 is Theorem 2 which tells us that a Gibbs Markov chain will be positive recurrent if and only if the full conditionals correspond to a proper posterior. Just after presenting this, Casella goes on to show us (7), which at first glance suggests that useful information cannot be extracted from Markov chains which are not positive recurrent.

I believe that such a conclusion is incorrect. To see why, I would like to discuss some examples where lower dimensional positive recurrent components can easily be extracted from Markov chains which are not positive recurrent.

The simplest and most obvious such example is obtained by interleaving a positive recurrent Markov chain $\Phi^1 = \Phi_1^1, \Phi_2^1, \ldots$, with a non positive recurrent Markov chain $\Phi^2 = \Phi_1^2, \Phi_2^2, \ldots$, to obtain $\Phi \equiv (\Phi_1^1, \Phi_1^2), (\Phi_2^1, \Phi_2^2), \ldots$ which is clearly not positive recurrent. Trivially, information in Φ about Φ^1 can be exploited by simply ignoring the Φ^2 components. Note that (7) does not apply to such functions because the conditions on t require that it be arbitrarily small outside of a compact set. This rules out functions which ignore the Φ^2 components, since these cannot be controlled over the range of Φ^2.

Based on this example, it may be tempting to think that the independence of Φ^1 and Φ^2 is what allows us to extract the positive recurrent chain. However, independence is not needed, as is illustrated by the following two examples.

In the first example, suppose the Gibbs sampler is used to generate a Gibbs chain $(x_1, y_1), (x_2, y_2), \ldots$ from the full conditionals

$$f_1(x|y) \propto e^{-(x+y)^2/2} \quad \text{and} \quad f_2(y|x) \propto e^{-(x+y)^2/2}. \tag{1}$$

The conditionals f_1 and f_2 are only functionally compatible, corresponding to an improper joint density of the form $f(x, y) \propto e^{-(x+y)^2/2}$. Thus, by Theorem 2, the Gibbs chain cannot be positive recurrent. Indeed, the subsequences $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ are interrelated random walks. This can be seen by noting that the Gibbs chain is obtained by successive substitution into

$$x_i = -y_{i-1} + \epsilon_i^x \quad \text{and} \quad y_i = -x_i + \epsilon_i^y \tag{2}$$

where ϵ_i^x and ϵ_i^y are independent $N(0, 1)$ variables. However, it is also clear from this representation that the derived Markov chain $z_1, z_2, \ldots$ where $z_i \equiv x_i + y_i = \epsilon_i^y$ is simply an $iid\ N(0, 1)$ sequence, obviously positive recurrent.

The second example is the one from Casella and George (1992) where the Gibbs sampler is used to generate a Gibbs chain from the full conditionals

$$f_1(x|y) \propto ye^{-xy} \quad \text{and} \quad f_2(y|x) \propto xe^{-xy}. \tag{3}$$

As Casella points out the conditionals f_1 and f_2 are only functionally compatible, corresponding to an improper joint density $f(x, y) \propto e^{-xy}$. Here too, the Gibbs chain cannot be positive recurrent. However, here the Gibbs chain $(x_1, y_1), (x_2, y_2), \ldots$ is obtained by successive substitution into

$$x_i = \epsilon_i^x / y_{i-1} \quad \text{and} \quad y_i = \epsilon_i^y / x_i \tag{4}$$

where ϵ_i^x and ϵ_i^y are independent exponential variables with mean 1. Thus, the derived Markov chain $z_1, z_2, \ldots$ where $z_i \equiv x_i y_i = \epsilon_i^y$ is simply an iid exponential sequence, again positive recurrent.

In both of the above examples, a positive recurrent chain $z_1, z_2, \ldots$ was constructed from the non positive recurrent chain (x_1, y_1), $(x_2, y_2), \ldots$. It is interesting to consider how the distribution of z arises through formal transformation of the improper density $f(x, y)$ corresponding to the Gibbs conditionals. In the first example, where $f(x, y) \propto e^{-(x+y)^2/2}$, the joint distribution of $z = x + y$ and $w = y$ is obtained as $f(z, w) \propto e^{-z^2/2}$. In the second example, where $f(x, y) \propto e^{-xy}$, the joint distribution of $z = xy$ and $w = y$ is obtained as $f(z, w) \propto \frac{1}{w} e^{-z}$. In both of these examples, an improper joint distribution has been transformed into the product of a proper distribution on z and an improper distribution on w. Thus, in both of these examples $f(x, y)$ contains a proper one-dimensional component which can be extracted from the output of a Gibbs sampler.

In light of these examples, I would like to ask Casella about the Gibbs subsequence of overall means $\beta^{(j)}, j \geq 1$ from Example 4 where $a = b = 0$. When (if ever) is this subsequence a positive recurrent component of the Gibbs chain? I have a hunch that it will be positive recurrent when $\pi(\beta|y)$, the posterior of β, is proper, in which case the subsequence will converge to $\pi(\beta|y)$. Can this be checked for the Gibbs output from Example 4?

JUN S. LIU *(Stanford University, USA)*

Professor Casella has provided us with a timely exposition of an important aspect of modern Monte Carlo methods. Stimulated by this reading, I would like to take the liberty of bringing up a few ideas on two interesting issues.

Rao-Blackwellizing an Importance Sampler. Consider an importance sampling scheme for a two-component random vector. Following no-

tations of Professor Casella, we let the target distribution of (X, Y) be $f(x, y)$ and let the trial sampling distribution be $g(x, y)$. Of interest is the estimation of, say, $\tau = E_f\{h(X, Y)\}$, for a given integrable function h. This can be achieved by using either rejection sampling, as demonstrated by Professor Casella, or importance sampling (IS). Suppose that we have drawn samples $(x_1, y_1), \ldots, (x_n, y_n)$ from $g(x, y)$. A standard IS estimate of τ is

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y_i)h(x_i, y_i), \quad \text{where} \quad w(x, y) = \frac{f(x, y)}{g(x, y)}.$$

A rescaled estimate, as illustrated in Section 4.2 and used in Casella and Robert (1996b), Kong et al. (1994), Liu (1996) etc., is

$$\tilde{\tau} = \frac{1}{W}\sum_{i=1}^{n} w(x_i, y_i)h(x_i, y_i), \quad \text{where} \quad W = \sum_{i=1}^{n} w(x_i, y_i).$$

Besides the advantage mentioned by Professor Casella, using the rescaled estimate $\tilde{\tau}$ allows us the flexibility of knowing f and g only up to a normalizing constant. This advantage is much more pronounced in complicated problems (Kong et al. 1994). Because asymptotically the two estimates are equivalent and also because $\hat{\tau}$ is much more approachable mathematically, we will use $\hat{\tau}$ for theoretical discussions, although practically we advocate using $\tilde{\tau}$ all the time.

There are two ways of Rao-Blackwellizing: conditioning on either X or Y. If conditioned on Y, for example, we have

$$\begin{aligned} E_g\{w(X, Y)h(X, Y) \mid Y = y\} &= \int h(x, y)\frac{f(x, y)}{g(x, y)}g(x \mid y)dx \\ &= w_y(y)E_f\{h(X, Y) \mid Y = y\}, \end{aligned}$$

where $w_y(y) = f_y(y)/g_y(y)$. A more efficient estimate than $\hat{\tau}$ results:

$$\hat{\tau}_{rby} = \frac{1}{n}\sum_{i=1}^{n} w_y(y_i)E_f\{h(X, Y)) \mid Y = y_i\}.$$

When h is a function of one component alone, say $h(x, y) = h(y)$, the estimate $\hat{\tau}_{rby}$ is reduced to

$$\hat{\tau}_{rby} = \frac{1}{n}\sum_{i=1}^{n} w_y(y_i)h(y_i).$$

A quite different intuitive interpretation of this R-B effect is that *marginalization* reduces importance sampling variation. MacEachern, Clyde, and Liu (1996) derived one special case of this fact, and Rubinstein (1981, Section 4.3.7) recorded another.

Under this formulation, the importance sampling can be treated approximately as a Rao-Blackwellized rejection sampling; hence, it is statistically more efficient. This fact has been established by Casella and Robert (1996b) in a sophisticated setting and will be re-derived here more directly and heuristically. Let (I_i, y_i), $i = 1, \ldots, n$, be jointly drawn according to the acceptance-rejection rule; that is, the y_i are iid from a trial distribution $g(y)$, and the conditional distribution $[I_i \mid y_i]$ is Bernoulli$(r(y_i))$ with $r(y) = f(y)/Mg(y)$. Suppose the stopping effect of this rejection sampling can be safely ignored. Then I_i plays the role of x_i in the foregoing argument; and the R-B counterpart of δ_{AR} in (10) of Casella is

$$\delta_{IS} = \frac{1}{n}\sum_{i=1}^{n} w(y_i)h(y_i).$$

Without loss of generality we assume that $\tau = 0$. Then, since $M \geq \max_y\{w(y)\}$,

$$\begin{aligned} n\text{var}(\delta_{AR}) &\approx M\text{var}_f\{h(Y)\} \geq \int w_{\max} h^2(y) f(y) dy \\ &\geq \int \frac{f(y)}{g(y)} h^2(y) f(y) dy = E_g\{w^2(y)h^2(y)\} \\ &= \text{var}_g\{w(y)h(y)\} = n\text{var}(\delta_{IS}). \end{aligned}$$

An effort of comparing the two samplers with the Metropolized independence sampling was made in Liu (1996). Since the advantage of the rejection method is that exact draws from f can be obtained, it is sometimes useful to combine the two samplers when one wants to reduce importance sampling variations (Liu, Chen, and Wong 1996).

In many practical problems, the marginal weight $w_y(y)$ is difficult to compute, whereas the conditional expectation $E_f\{h(X) \mid Y = y\}$ is relatively easy to obtain. In such cases, as shown in Kong et al. (1994), one can use a partial RB-estimate

$$\hat{\tau}_{prb} = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y_i) E_f\{h(X, Y) \mid Y = y_i\},$$

which is easily seen to be unbiased and consistent. Although many numerical results show that significant improvements can be obtained, optimality properties of $\hat{\tau}_{prb}$ are difficult to come by.

Imagine that a partial R-B is applied twice; then each summand of $\hat{\tau}_{prb}$, $E_f\{h(X,Y) \mid Y = y_i\}$, is substituted by $E_f[E_f\{h(X,Y)|Y\}| X = x_i]$. By applying partial R-B repeatedly, each summand has the form of iterative conditional expectations:

$$E_f[\cdots E_f\{E_f\{h(X,Y) \mid Y\} \mid X\} \cdots \mid \cdot],$$

whose limit converges to the true value τ. This form alludes to the Gibbs sampling structure (Liu, Wong and Kong 1994, 1995). When analytical evaluation of these iterative conditional expectations is not feasible, one is naturally reminded of the Gibbs sampler. A suggestion thus derived is that incorporating a Gibbs sampler or any MCMC step into an importance sampling scheme can be useful (MacEachern et al. 1996).

The Gibbs Sampler for Incompatible Conditionals

An impressive result of Hobert and Casella (1996) is concerned with the stochastic instability of Gibbs sampling with incompatible — but functionally compatible — conditionals. I would like to venture on the functionally incompatible case. Consider the following example: suppose that the two conditionals $f_1(y|x)$ and $f_2(x|y)$ are given as follows:

$f_1(y|x)$:

	$y=1$	$y=2$
$x=1$	0.9	0.1
$x=2$	0.3	0.7

$f_2(x|y)$:

	$x=1$	$x=2$
$y=1$	0.4	0.6
$y=2$	0.2	0.8

It is easy to show that f_1 and f_2 are not functionally compatible using Besag's (1974) criterion. When running a systematic-scan Gibbs sampler, the concept of "limiting distribution" becomes a little complicated. In fact, the sampler has two limiting distributions depending on whether stopping at x or at y, i.e., whether (x,y) or (y,x) is defined as a joint state. The two limiting distributions are

$\pi_1(x,y)$:

	$y=1$	$y=2$
$x=1$	0.26591	0.02955
$x=2$	0.21136	0.49318

$$\pi_2(x,y): \begin{array}{c|cc} & y=1 & y=2 \\ \hline x=1 & 0.19091 & 0.10455 \\ x=2 & 0.28636 & 0.41818 \end{array}$$

The sampler is, therefore, a combination of two positive recurrent Markov chains; and depending on how to define the joint state, the sampler converges into two different, though very close, distributions. When running a random-scan Gibbs sampler, however, a proper limiting distribution — that is the mixture of the two distributions given above — exists.

Under some regularity conditions that are satisfied in most practical situations, $T_x(x_0, x_1) = \int f_1(y|x_0) f_2(x_1|y) dy$ defines a positive recurrent transition function for the X space, and $T_y(y_0, y_1) = \int f_2(x|y_0) f_1(y_1|x) dx$ defines that for the Y space. Hence two limiting distributions $\pi_1(x)$ and $\pi_2(y)$, for T_x and T_y, respectively, are uniquely determined. In the incompatible case, we observe that

$$\pi_1(x,y) \equiv \pi_1(x) f_1(y \mid x) \neq \pi_2(y) f_2(x \mid y) \equiv \pi_2(x,y).$$

But

$$\int \pi_1(x) f_1(y \mid x) dx = \pi_2(y) \quad \text{and} \quad \int \pi_2(y) f_2(x \mid y) dy = \pi_1(x).$$

Let $\mathcal{P}_1$ be the set of all probability distributions compatible with $f_1(y|x)$, and let $\mathcal{P}_2$ be that for $f_2(x|y)$. Then $\pi_1(x,y) \in \mathcal{P}_1$, $\pi_2(x,y) \in \mathcal{P}_2$, and π_1 and π_2 have identical marginal distributions. On the other hand, if two distributions $p_1(x,y) \in \mathcal{P}_1$ and $p_2(x,y) \in \mathcal{P}_2$ have identical marginal distributions, they have to be the same as π_1 and π_2.

Due to numerical approximation in practice, we may end up having slightly incompatible conditionals. If the numerical error is small, the resulting T_x will be very close to the one, say, T_x^*, resulting from the compatible conditionals. This implies that the eigenvalues and eigenvectors of T_x and T_x^* are close to each other (true in the finite state space case); hence, the resulting limiting distributions are similar. It further suggests that no disasters are to be expected as long as the numerical approximation is reasonably accurate. The argument may be extended to a Gibbs sampler with more than two components. For a k component sampler, a systematic scan with a particular sweeping order will have

k limiting distributions, depending on which component the sampler stops. The total number of such limiting distributions is $k!$. The limiting distribution for a random-scan sampler is then a mixture of these $k!$ distributions.

XIAO-LI MENG *(The University of Chicago, USA)*

Posterior Checking. My discussion will focus on only one issue: checking the propriety of a posterior resulting from the Gibbs-sampler specifications. Professor Casella's article is much broader, touching on many issues that are of current interest to me (e.g., the emphasis on being receptive to both frequentist and Bayesian perspectives; the interplay of algorithms and inferences; the connection between EM-type algorithms and the Gibbs sampler). However, due to stringent time constraints (being a father of a newborn and a 16-month-old, I had to prepare this discussion in between frequent posterior checking; no impropriety was found, though I did learn why it is a good idea to avoid a sensitive posterior), I have to skip this great opportunity for advertising several related papers that I authored or co-authored. Nevertheless, I want to thank the Editor, and of course the author, for providing me with such an opportunity.

Recursive De-conditioning and Conditional Compatibility. The need for checking the compatibility of conditional distributions reminds me of an identity I learned more than a year ago. Let $p(x_1, x_2)$ be a probability density function with respect to a product measure $\mu = \mu_1 \times \mu_2$ and with a support in the form $\Omega_1 \times \Omega_2$; we thus are assuming the *positivity* assumption of Hammersley and Clifford (c.f., Besag, 1974). Then

$$p(x_1) = \left[\int_{\Omega_2} \frac{p(x_2 \mid x_1)}{p(x_1 \mid x_2)} \mu_2(dx_2) \right]^{-1}, \tag{1}$$

which is a trivial consequence of the well-known identity

$$\frac{p(x_2 \mid x_1)}{p(x_1 \mid x_2)} = \frac{p(x_2)}{p(x_1)}. \tag{2}$$

While identity (1) also provides an explicit formula showing how $p(x_1 \mid x_2)$ and $p(x_2 \mid x_1)$ uniquely determine $p(x_1, x_2)$, it seems to be much

less well-known than the standard formula for proving uniqueness:

$$p(x_1) \propto \frac{p(x_1 \mid x_2')}{p(x_2' \mid x_1)}, \quad \text{for any fixed } x_2' \in \Omega_2, \tag{3}$$

which also is an immediate consequence of (2).

I learned the expression (1) from a presentation by Ng (1995). My immediate reaction was that it must be my ignorance that I had not seen (1) in this explicit form. However, Ng assured me that he had checked with several leading experts in this area (e.g. J. Besag, W. H. Wong), and it seemed that the identity (1) was "mysteriously" missing from the general literature. An apparent explanation for this "mystery" is that (1) is not useful in general for calculating $p(x_1)$ and thus $p(x_1, x_2)$ since a main reason we use the Gibbs sampler is our inability to perform analytical integration, which is required by (1). However, in the context of checking the compatibility of $p(x_1 \mid x_2)$ and $p(x_2 \mid x_1)$, the expression (3) offers no advantage over (1). Both require us first to check whether $p(x_1 \mid x_2)$ and $p(x_2 \mid x_1)$ are functionally compatible, which amounts to checking whether (2) is possible, that is, whether we can write

$$\frac{p(x_2 \mid x_1)}{p(x_1 \mid x_2)} = \frac{\tilde{p}_2(x_2)}{\tilde{p}_1(x_1)} \tag{4}$$

for some (positive) functions $\tilde{p}_i$, $i = 1, 2$. Given (4) holds, we then need to check, for (1), whether $\int_{\Omega_2} \tilde{p}_2(x_2)\, \mu_2(dx_2)$ is finite, or, for (3), whether $\int_{\Omega_1} \tilde{p}_1(x_1)\, \mu_1(dx_1)$ is finite. Under (4), these two integrations must yield the same value (allowing $+\infty$) by Fubini's theorem, and thus one can always choose one to check (e.g., x_1 and x_2 may be of very different dimensions), as emphasized by Arnold and Press (1989). Of course, these arguments also imply that there is no advantage to using (1) in the simple case involving only $p(x_1 \mid x_2)$ and $p(x_2 \mid x_1)$.

Reading Section 3 of Casella's article (and Hobert and Casella, 1995) made me wonder about the comparison between (1) and (3) for checking the compatibility of $\{p(x_i \mid X_{-\{i\}}), 1 \le i \le m\}$ when $m > 2$, where $X = \{x_1, \ldots, x_m\}$ and X_{-S} denotes $\{x_j, j \notin S\}$. I thus decided to take a closer look at this comparison and the rest of this discussion reports what it generated. I doubt anything I discuss here is new (though I have not seen the recursive scheme described below), since everything follows

in a straightforward manner from (2); my discussion is thus more of a review nature, intended as a technical supplement to Casella's general review of the important issue of checking compatibility.

For $m > 2$, a direct generalization of (3) is (see Besag, 1974; Gelman and Speed, 1993; Hobert and Casella, 1995)

$$p(x_2, \ldots, x_m) \propto \frac{\prod_{j=2}^{m} p(x_j \mid x_1, x_2, \ldots, x_{j-1}, x'_{j+1}, \ldots, x'_m)}{\prod_{j=2}^{m} p(x'_j \mid x_1, x_2, \ldots, x_{j-1}, x'_{j+1}, \ldots, x'_m)}, \quad \text{for any fixed } (x'_2, \ldots, x'_m) \in \prod_{k \geq 2} \Omega_k. \tag{5}$$

Since the indices $(1, \ldots, m)$ are arbitrary, we actually have $m!$ ways of obtaining $p(x_1, \ldots, x_m)$ via (5). Specifically, Hobert and Casella (1995) define

$$g_i(x_1, \ldots, x_m) = \frac{\prod_{j=1}^{m} p(x_{l_j^i} \mid x_{l_1^i}, \ldots, x_{l_{j-1}^i}, x'_{l_{j+1}^i}, \ldots, x'_{l_m^i})}{\prod_{j=2}^{m} p(x'_{l_j^i} \mid x_{l_1^i}, x_{l_2^i}, \ldots, x_{l_{j-1}^i}, x'_{l_{j+1}^i}, \ldots, x'_{l_m^i})}, \tag{6}$$

where $l^i = (l_1^i, l_2^i, \ldots, l_m^i)$ represents a permutation of $(1, \ldots, m)$ and $(x'_1, \ldots, x'_m)$ is a fixed point in $\Omega \stackrel{def}{=} \Omega_1 \times \cdots \times \Omega_m$. Hobert and Casella (1995) then show that $\{p(x_i \mid X_{-\{i\}}),\ i = 1, \ldots, m\}$ are functionally compatible if and only if there is a (positive) function $g(x_1, \ldots, x_m)$ on Ω such that $g_i(x_1, \ldots, x_m) \propto g(x_1, \ldots, x_m)$. Furthermore, if $\{p(x_i \mid X_{-\{i\}}),\ i = 1, \ldots, m\}$ are functionally compatible, then they are compatible if and only if

$$\int_{\Omega_1} \cdots \int_{\Omega_m} g(x_1, \ldots, x_m)\, \mu_m(dx_m) \cdots \mu_1(dx_1) < \infty. \tag{7}$$

Finally, $p(x_1, \ldots, x_m) \propto g(x_1, \ldots, x_m)$ when (7) holds.

To apply (1) for $m > 2$, we first note a conditional version of (1), that is, for any $A \supset \{i, j\},\ i \neq j$,

$$p(x_i \mid X_{-A}) = \left[\int_{\Omega_j} \frac{p(x_j \mid x_i, X_{-A})}{p(x_i \mid x_j, X_{-A})} \mu_j(d\, x_j) \right]^{-1}. \tag{8}$$

The right-hand side of (8) may be viewed as a "de-conditioning" operator, that is, with the help of $p(x_j \mid x_i, X_{-A})$, it turns $p(x_i \mid x_j, X_{-A})$ into $p(x_i \mid X_{-A})$ — de-conditioning out x_j. It is obvious that this de-conditioning operator can be applied recursively to further de-condition out variables in X_{-A}. To be more precise, let $\mathcal{F}$ be the set of positive functions (allowing the value $+\infty$) on Ω (almost surely with respect to $\mu \stackrel{def}{=} \mu_1 \times \cdots \times \mu_m$; hereafter, I will not repeat such measure-theoretic statements). For any $1 \leq k \leq m$, we define a mapping $\mathcal{D}_k$ from $\mathcal{F} \times \mathcal{F}$ to $\mathcal{F}$, such that for any $f_1, f_2 \in \mathcal{F}$:

$$\mathcal{D}_k[f_1 : f_2] = \left[\int_{\Omega_k} \frac{f_1(x_1, \ldots, x_k, \ldots, x_m)}{f_2(x_1, \ldots, x_k, \ldots, x_m)} \mu_k(dx_k) \right]^{-1}. \quad (9)$$

Now for a given set of conditionals $\{p(x_i \mid X_{-\{i\}}),\ i = 1, \ldots, m\}$, we view them as elements of $\mathcal{F}$ and label $f_{i1} = p(x_i \mid X_{-\{i\}})$, $i = 1, \ldots, m$. We then define $\{f_{ij},\ i = j, \ldots, m;\ j = 2, \ldots, m\}$ recursively via

$$f_{ij} = \mathcal{D}_{j-1}[f_{j-1,j-1} : f_{i,j-1}], \qquad i = j, \ldots, m;\ j = 2, \ldots, m. \quad (10)$$

Clearly, f_{ij} depends on X only through $\{x_j, \ldots, x_m\}$ so we write $f_{ij}(x_j, \ldots, x_m)$ whenever explicit arguments are needed. By (8), it is easy to show via induction that if $\{f_{i1},\ i = 1, \ldots, m\}$ are derived from a joint density $p(x_1, \ldots, x_m)$, then

$$f_{ij}(x_j, \ldots, x_m) = p(x_i \mid X_{-\{1,\ldots,j-1,i\}}), \quad \text{for any } i \geq j,\ j \geq 2, \quad (11)$$

and in particular

$$p(x_1, \ldots, x_m) = \prod_{j=1}^{m} f_{jj}(x_j, \ldots, x_m). \quad (12)$$

We thus learn that, in order to have compatibility of $\{f_{i1},\ i = 1, \ldots, m\}$, it is necessary that for any $m - 1 \geq j \geq 1$ and $i \geq j + 1$:

(I) f_{jj} and f_{ij} are functionally compatible *conditional* on $X_{-A_{ij}}$ where $A_{ij} = \{1, \ldots, j, i\}$; namely, we can find functions $\tilde{f}_i(x_i; X_{-A_{ij}})$

$\in \mathcal{F}$ and $\tilde{f}_j(x_j; X_{-A_{ij}}) \in \mathcal{F}$ such that

$$\frac{f_{jj}(x_j, \ldots, x_m)}{f_{ij}(x_j, \ldots, x_m)} = \frac{\tilde{f}_j(x_j; X_{-A_{ij}})}{\tilde{f}_i(x_i; X_{-A_{ij}})}, \quad \text{for } (x_j, \ldots, x_m) \in \prod_{k \geq j} \Omega_k; \tag{13}$$

and

(II) The functions $\tilde{f}_i$ and $\tilde{f}_j$ found in (13) must satisfy

$$\int \tilde{f}_j(x_j; X_{-A_{ij}})\, \mu_j(dx_j) = \int \tilde{f}_i(x_i; X_{-A_{ij}})\, \mu_i(dx_i) < +\infty, \quad \text{for } X_{-A_{ij}} \in \prod_{k>j, k\neq i} \Omega_k. \tag{14}$$

Conditions (I) and (II) amount to the *conditional compatibility* of f_{jj} and f_{ij} conditional on $X_{-A_{ij}}$. Because of (12), these conditions are also sufficient for the compatibility of $\{f_{i1},\ i = 1, \ldots, m\}$. In other words, $\{p(x_i \mid X_{-\{i\}}),\ i = 1, \ldots, m\}$ are compatible if and only if (I) and (II) are satisfied for all $m - 1 \geq j \geq 1$ and $i \geq j + 1$.

A matrix representation of $\{f_{ij},\ i \geq j, j = 1, \ldots, m\}$ perhaps can help to visualize the recursive de-conditioning process defined by (10). Table 1 gives the representation with $m = 4$, where we use $[\cdot \mid \cdot]$ to denote conditional density (e.g., $[4|3] \stackrel{def}{=} p(x_4|x_3)$) and "$k$" to indicate the elimination (i.e.,"de-conditioning") of x_k from the variables that are being conditioned on.

Table 1. *A Matrix Representation of Recursive De-conditioning*

f_{ij}	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$i = 1$	$[1 \mid 234]$			
$i = 2$	$[2 \mid 134]$	$[2 \mid \not{1}34] \equiv [2 \mid 34]$		
$i = 3$	$[3 \mid 124]$	$[3 \mid \not{1}24] \equiv [3 \mid 24]$	$[3 \mid \not{2}4] \equiv [3 \mid 4]$	
$i = 4$	$[4 \mid 123]$	$[4 \mid \not{1}23] \equiv [4 \mid 23]$	$[4 \mid \not{2}3] \equiv [4 \mid 3]$	$[4 \mid \not{3}] \equiv [4]$

The matrix representation makes it easier to track the de-conditioning process, especially because each column corresponds to de-conditioning out one variable, starting from the finest conditioning ($j = 1$) recursively down to no conditioning ($j = m$). It also makes it clear that $\{f_{i1},\ i = 1, \ldots, m\}$ are compatible if and only if $\{f_{ij},\ i \geq j\}$ are *conditionally compatible* (as defined by (I) and (II)) for each $j = 1, \ldots, m-1$.

To illustrate the use of (I) and (II) for checking compatibility, let us consider the normal example used by Hobert and Casella (1995):

$$f_{i1} \equiv p(x_i \mid X_{-\{i\}}) \propto \exp\{-\frac{1}{2}(x_i - p_i \sum_{k \neq i}^{m} x_k)^2\}, \quad i = 1, \ldots, m. \tag{15}$$

Here the p_i's are constants, and the goal is to identify conditions on p_i's under which $\{p(x_i \mid X_{-\{i\}}),\ i = 1, \ldots, m\}$ are compatible. Since for any $i > 1$ the only term in the exponential part of f_{11}/f_{i1} involving $x_1 x_i$ is $(p_1 - p_i)x_1 x_i$, (13) is satisfied if and only if $p_1 = p_i$. This yields a necessary condition for the compatibility: $p_i \equiv p$ for all i. Under this necessary condition,

$$\frac{f_{11}}{f_{i1}} = \frac{\exp\left\{-\frac{(1-p^2)}{2}\left(x_1 - \frac{p}{1-p}T_{1i}\right)^2\right\}}{\exp\left\{-\frac{(1-p^2)}{2}\left(x_i - \frac{p}{1-p}T_{1i}\right)^2\right\}}, \tag{16}$$

where $T_{1i} = \sum_{k \notin \{1,i\}} x_k$. It then follows that (14) holds if and only if $p^2 < 1$, under which

$$f_{i2} \propto \exp\left\{-\frac{(1-p^2)}{2}\left(x_i - \frac{p}{1-p}T_{1i}\right)^2\right\}, \quad i = 2, \ldots, m. \tag{17}$$

No further integration is needed if we notice that checking the conditional compatibility of (17) is the same as that of (15) with $p_i \equiv p$, in the sense that both can be written as

$$f_{ij} \propto \exp\left\{-\frac{c_j}{2}(x_i - \beta_j \sum_{k \geq j,\ k \neq i} x_k)^2\right\}, \quad i = j, \ldots, m, \quad j = 1, 2, \tag{18}$$

where $c_1 = 1, c_2 = 1 - p^2, \beta_1 = p, \beta_2 = \beta_1/(1 - \beta_1) = p/(1 - p)$. Thus $\{f_{i2},\ i = 2, \ldots, m\}$ are conditionally compatible if and only if $\beta_2^2 < 1$. By induction, for $j = 3, \ldots, m - 1$, $\{f_{ij},\ i = j, \ldots, m\}$ are conditionally compatible if and only if $\beta_j^2 < 1$, where $\beta_j = \beta_{j-1}/(1 - \beta_{j-1}) = p/(1 - (j - 1)p)$. Thus $\{f_{i1},\ i = 1, \ldots, m\}$ are compatible if and only if $\beta_j^2 < 1$ for all $j = 2, \ldots, m - 1$, which is equivalent to $-1 < p < 1/(m - 1)$. Hobert and Casella (1995) used (5)-(7) to reach this conclusion, which can also be obtained by noticing that the common correlation among $\{x_1, \cdots, x_m\}$ is given by $p/(1 - (m - 2)p)$, which must be between $-1/(m - 1)$ and 1, exclusively.

Of course, the simplicity of this example is largely due to the simplicity of the model, especially due to the normality which is preserved under de-conditioning. In general, the requirement of analytically calculating the $\mathcal{D}_k$ mapping contradicts the goal of using the Gibbs sampler, and thus the recursive de-conditioning method via the $\mathcal{D}_k$ mapping, when used as a *sufficient* check, is typically useless in practice when $m > 2$ (except for special conditional densities, such as normal). This perhaps further explains why this method, though mathematically interesting, has been ignored in the literature (except, perhaps, in the written version of Ng (1995), which I have not had an opportunity to study).

Fortunately, the comparative study is not without any positive message. The recursive de-conditioning scheme itself, as depicted in Table 1, has something to be recommended. In contrast to (5)-(7), it involves only two (conditional) functions at a time, and the check of the integrability only involves marginal integrations (see (14)). More importantly, it can tell us at which level of conditioning the densities (in fact which conditional density) become improper (e.g., for the normal example, $\{p(x_i \mid X_{-\{1,\ldots,j-1,i\}}), i \geq j\}$ are proper for all $j \leq k$ but are improper when $j = k + 1$ if and only if $(k - 1)^{-1} > p \geq k^{-1}$, where $2 \leq k \leq m - 1$). Such specific information can be useful when we modify parts of the model in order to achieve compatibility. In particular, the conditional compatibility at the $j = 1$ level (see Table 1) can and should be checked first, since such a check does not require explicit calculation of the $\mathcal{D}$ mapping and if the conditional compatibility is violated (e.g., if some of f_{i2}'s are determined to be improper) then our check is completed. (For the normal example, such a check immediately declares that if any $p_i \neq p_j$, or if the common $p^2 \geq 1$, then the

conditional distributions given in (15) are incompatible.) As a *necessary* check (i.e., a screening check), this can be considerably simpler than the check using (6)-(7), which operates on the entire joint space. In some cases, it might even be possible to continue this check for conditional compatibility for a few more levels (e.g., $j = 2$ or 3) if we can arrange the variables $x_1, \ldots, x_m$ such that the first few $\mathcal{D}_j$ mappings are analytically feasible. It is also not entirely inconceivable that we can check the integrability of ratios of $\mathcal{D}_j$'s without explicitly calculating $\mathcal{D}_j$.

Of course, ideally we would like to have a recursive de-conditioning scheme, similar to Table 1, using mappings that do not involve integration. For example, it would be ideal if we could use the mapping defined by the following conditional version of (3):

$$p(x_1 \mid X_{-A}) \propto \frac{p(x_1 \mid x_2', X_{-A})}{p(x_2' \mid x_1, X_{-A})}, \quad \text{for any } A \supset \{1,2\} \text{ and any fixed } x_2' \in \Omega_2. \tag{19}$$

Although (19) is true, it does not yield a correct de-conditioning process when used recursively in a fashion similar to (10) because the normalizing constant in (19) depends on A. I suspect that it is impossible to perform the type of recursive de-conditioning depicted in Table 1 without invoking integration (i.e., marginalization). However, it might be possible to construct a recursive checking scheme that is more effective than the check based on (6)-(7), which is essentially a brute-force method and can be rather complicated (see, e.g., Hobert and Casella's (1996) proof of the quoted Theorem 1). I know Professor Casella enjoys working on challenging theoretical constructions, so I'd like to conclude my discussion by inviting him to a fishing trip for an effective recursive checking scheme. I cannot promise we will get anything, but the excitement of fishing (my favorite sport) is not knowing what you will get or when you will get it — there is always a bigger one out there, the one that snapped my line before I could see it!

Acknowledgments. I thank K. Ng for an informative presentation, and A. Gelman, C. Liu, W. Rosenberger, and A. Zaslavsky for comments. The research was supported in part by NSA Grant MDA 904-96-1-0007 and NSF Grant DMS-9626691. This manuscript was prepared using computer facilities supported in part by several NSF grants awarded to

the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

A. PHILIPPE *(Université de Rouen, France)*

I first want to congratulate Professor Casella on such a coverage of the multiple facets of the relationship between statistical theory and computational algorithms. I want to take advantage of this tribune to point out links between the Monte Carlo method with numerical methods used to approximate integrals. The standard Monte Carlo estimator is the empirical average. The convergence of this type of estimator is ensured by the Law of Large Numbers or the ergodic theorem. In this paper Professor Casella looks at the amount of statistical theory in the Monte Carlo method. The outputs of the Monte Carlo algorithm are considered as statistical data and therefore we can apply frequentist principles to improve upon the standard approach. An alternative to this approach is to consider the output as a set of points on which we can apply numerical quadrature. In particular, when we generate a sample from a density f, we can use it to build a Riemann sum, i.e. the trapezoidal approximation of the integral.

This method has been introduced by Yakowitz *et al* (1978) in the particular case of the uniform distribution, i.e. for functions with compact support. They show that the estimator thus produced improves (in terms of convergence rate) upon the empirical average as it reduces its variance. The properties obtained for this particular density can be generalized for arbitrary densities f (Philippe 1996). We discuss the different aspects of using Riemann sums in the Monte Carlo method. In the case of the Gibbs sampler, we show that we can produce an efficient estimator based on the Rao-Blackwellisation method and Riemann sums.

1. Riemann sums and the Monte Carlo method. Consider the estimation of the expectation $\mathbb{E}^f[h]$, where f is a density and $h \in \mathcal{L}^1(f)$ is a continuous function. For a sample $(x_1, \cdots, x_n)$ from f, we denote the ordered sample by $x_{(1)} \leq \ldots \leq x_{(n)}$. The resulting estimator (called Riemann's estimator) is given by

$$\delta_n^R = \sum_{i=1}^{n-1} \left(x_{(i+1)} - x_{(i)}\right) h\left(x_{(i)}\right) f\left(x_{(i)}\right). \tag{1.1}$$

The convergence properties of the Riemann estimator are given in the following propositions.

Proposition 1.1. *If* $h \in \mathcal{L}^1(f)$ *then*

$$\lim_{n\to\infty} \mathbb{E}\left[\delta_n^R\right] = \mathbb{E}^f[h].$$

Moreover if the function h *is bounded on the support of* f *then the convergence rate of the bias is* $O(n^{-1})$.

Proposition 1.2. *If* $h \in \mathcal{L}^2(f)$ *then*

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\delta_n^R - \mathbb{E}\left[\delta_n^R\right]\right)^2\right] = 0.$$

Moreover if h *and* h' *are bounded on the support of* f *then the convergence rate of the variance is* $O(n^{-2})$.

These convergence properties clearly show the improvement brought by this approach upon the standard Monte Carlo averaging approach. Indeed, when the previous conditions on h are satisfied, the behavior of the Riemann estimator is very satisfactory since it reduces the variance by an order of magnitude, that is, from $1/n$ to $1/n^2$. However, in many statistical problems, the function h is not bounded. For example, a classical problem, in Bayesian statistics, is the evaluation of the Bayes estimator. Under the quadratic loss, this is the mean of the posterior distribution, so $h(x) = x$ which is unbounded for infinite support.

An additional appeal of our approach is that the importance sampling method can improve upon the Riemann estimator, while keeping the same convergence properties for bounded h's. This improved Riemann estimator follows from the choice of an instrumental function g such that the ratio hf/g and its derivative are bounded. It is produced through $\mathbb{E}^f[h] = \mathbb{E}^g[hf/g]$ and equals to

$$\sum_{i=1}^{n-1} \left(y_{(i+1)} - y_{(i)}\right) h\left(y_{(i)}\right) f\left(y_{(i)}\right)$$

where $y_{(1)} \leq \cdots \leq y_{(n)}$ is an ordered sample of variables with density g. Note that the density does not appear explicitly in the expression of the estimator. A good choice of the instrumental function is a density proportional to $|h|f$. This choice is optimal in terms of reduction of the

variance when the support of the density is bounded. Furthermore it gives an unbiased estimator when the function h is positive.

This choice is also optimal for the standard importance sampling method (see Rubinstein 1981), although this result is formal. Indeed the estimator depends on the ratio f/g; therefore the unknown integral of interest appears in the expression of the estimator. The Riemann estimator based on the instrumental density proportional to $|h|f$ is easy to derive via an accept-reject algorithm. The only requirement is to find g such that the ratio $|h|f/g$ is bounded.

Example 1. Consider the example of the gamma distribution introduced by Professor Casella. The gamma distribution $\mathcal{G}a(\alpha, 2\alpha)$ with $\alpha = 2.434$ is simulated from an accept-reject algorithm where the candidate distribution is the gamma distribution $\mathcal{G}a(a, 2a)$ with $a = 2$. We want to estimate the expectation $\mathbb{E}^f(x)$. With the same instrumental density $\mathcal{G}a(a, 2a)$, we can also generate a sample from the density proportional to hf. Table 1.1 illustrates the behavior of the different Riemann estimators. We can appreciate the superior properties of the Riemann estimator obtained with the sample simulated from the density proportional to hf. Moreover, this estimator dominates the estimators produced by the Rao-Blackwell strategy, since the percent improvement in mean squared error (MSE) is superior for this Riemann estimator.

Table 1.1. *Comparison of the mean squared errors for the estimation of a gamma mean given by the empirical average, the Riemann estimators δ_1^R and δ_2^R obtained respectively with the sample simulated from $\mathcal{G}a(\alpha, 2\alpha)$ and from the density proportional to hf, based on* 7500 *simulations.*

AR sample size (t)	δ^E MSE	δ_1^R MSE	δ_2^R MSE	Pourcent Decrease in MSE for δ_2^R
25	.0041	.0060	.0021	48.78
50	.0020	.0026	.0006	70.00
100	.0010	.0009	.0001	90.00

Table 1.2. *Comparison of the mean squared errors for the estimation of a gamma mean given by the Riemann estimators recycling the N values produced by the accept-reject algorithm, for the sample from $\mathcal{G}a(\alpha, 2\alpha)$ ($\tilde{\delta}_1^R$) and for the sample from the density proportional to hf ($\tilde{\delta}_2^R$), based on* 7500 *simulations.*

AR sample size (t)	$\tilde{\delta}_1^R$ MSE	$\tilde{\delta}_2^R$ MSE
25	.0031	.002
50	.0012	.0002
100	.0004	.0001

For fixed t, the accept-reject algorithm generates $(y_1, \ldots, y_N)$ from the instrumental distribution and yields a sample $(x_1, \cdots, x_t)$ of size t from $\mathcal{G}a(\alpha, 2\alpha)$. The number of values N is a random integer which is distributed according to a geometric random variable. However, this sample can be interpreted as a sample simulated from the instrumental density $\mathcal{G}a(a, 2a)$, and therefore we construct the Riemann estimator from the sample $(y_1, \cdots, y_N)$ according to the importance sampling approach. This method recycles all the random variables produced by the accept-reject algorithm. We apply also this principle for the accept-reject algorithm which produces a sample from the density proportional to hf. Table 1.2 illustrates the behavior of the Riemann estimators. When we recycle the rejected variables, the performances of the Riemann estimators are superior since the mean squared errors is reduced.

2. The Rao-Blackwellisation method and the Riemann estimator.

An important problem with this form of estimators is that it requires explicit densities. However, in many statistical problems this condition is not satisfied (see for instance the Gibbs sampler) and (1.1) cannot be used. The Gibbs sampler method can generate a sample from f when the density is not directly available. It is indeed sufficient to know the conditional distributions. An alternative is to consider a modified form of the Riemann estimator by replacing the term which depends on f by an approximation. Note that this integral can also be considered as a

multiple integral. However, the generalization of the Riemann estimator to larger dimensions is not efficient, as shown by Yakowitz *et al.* (1978).

The Rao-Blackwellisation method produces an estimator of the marginal density (see Gelfand and Smith, 1990). This estimator of the density is given by

$$\hat{f}(x) = n^{-1} \sum_{t=1}^{n} \pi(x|x_2^t, \cdots, x_p^t). \tag{2.1}$$

Note that, when we use the Gibbs sampler algorithm, this estimator is available. Therefore, we can always get the following generalized form of the Riemann estimator :

$$\delta_n^{R/RB} = n^{-1} \sum_{t=1}^{n} (x_1^{(t+1)} - x_1^{(t)}) h(x_1^{(t)}) \left(\sum_{k=1}^{n} \pi(x_1^{(t)}|x_2^k, \cdots, x_p^k) \right). \tag{2.2}$$

The computational cost of this estimator is higher than for the standard Riemann estimator but the efficiency is quite similar and it definitely improves upon the empirical average. The performances are illustrated in the case of the auto exponential model (Besag, 1974).

Example 2. Consider the density

$$f(y_1, y_2) \propto \exp(-y_1 - y_2 - y_1 y_2).$$

The corresponding conditional distribution are given by

$$y_1|y_2 \sim \mathcal{E}xp(1 + y_2),$$
$$y_2|y_1 \sim \mathcal{E}xp(1 + y_1).$$

Since the marginal density is known up to a constant factor, i.e.

$$f_1(y_1) \propto \frac{e^{-y_1}}{1 + y_1},$$

we can compare the Riemann estimators (1.1) and (2.1) with the empirical average and the Rao-Blackwell estimator. By running a Monte-Carlo experiment 200 times, we build equal tailed confidence regions C_n such that, for fixed n,

$$P(\delta_n \in C_n) = 1 - \alpha.$$

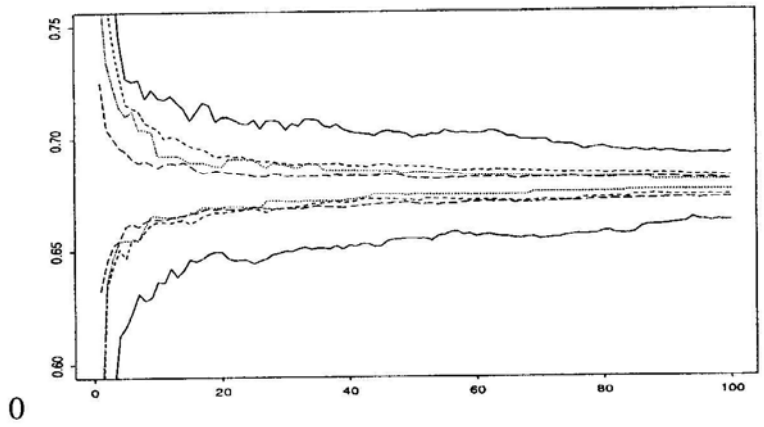

Figure 2.1. 95% *confidence band for the estimation of* $\mathbb{E}^f(x)$ *for the auto exponential model: the empirical average (plain), the Riemann estimator* (1.1) *(dots), the modified Riemann estimator* (2.2) *(dashes) and the Rao-Blackwell estimator (long dashes). For* $n = 5,000$, *the confidence band are* $[0.6627, 0.6932]$, $[0.6761, 0.6806]$, $[0.6738, 0.6825]$, *and* $[0.6728, 0.6807]$ *respectively and the true value is* 0.6768.

Figure 2.1 shows the behavior of the confidence band for $\alpha = 0.05$. The amplitude of the confidence band of the Riemann and Rao-Blackwell estimators are quite similar. The three estimators improve upon the empirical average.

JOSEPH L. SCHAFER *(The Pennsylvania State University, USA)*

I would like to thank Dr. Casella for a thoughtful and well-written paper. In this era of rapidly improving computer environments, many are tempted to adopt an algorithmic approach to inference. Monte Carlo (MC) methods—and Markov chain Monte Carlo (MCMC) in particular—have become a popular paradigm for statistical problem solving, but the results of MC or MCMC runs are only as good as (a) the underlying statistical model and (b) the manner in which the ouput stream is collected and summarized. Improvements to (b) are certainly worth considering; Casella and his colleagues have suggested some potentially useful methods. With regard to (a), of course, we should not expect MC to yield useful information if the underlying statistical model is nonsensical.

The methods of Sections 4–5 were motivated by principles of classical decision theory. A decision theoretic perspective can be helpful, provided that we pay attention to the MC simulation's original purpose. If the goal is to draw inferences about a parameter $h(\theta)$ of the data model for y, the Bayesian perspective suggests that we examine the posterior mean, variance, quantiles, etc. of $h(\theta)$. MC algorithms yield estimates of these quantities which can, in principle, be made as accurate as desired by lengthening the simulation run. Casella *et al.* focus on improving the efficiency of these MC estimates. That goal, however, is one step removed from the statistician's ultimate purpose. Any reasonable MC estimator of $E(h(\theta) \mid y)$, even if it is not highly efficient, will be good enough if its mean-squared error is small relative to $V(h(\theta) \mid y)$. Improving the efficiency of MC estimators is not necessarily profitable if it does not substantially improve the quality of the point and interval estimates for $h(\theta)$ itself.

A major theme of this paper is the interplay between the data model and the MC simulation method. I prefer to view the MC simulation an additional step of data collection, much like a second stage of sampling in a multistage survey. Let $S^{(m)}$ denote the output stream from a simulation run of length m. If computational resources were unlimited, we could generate $S^{(\infty)}$ and obtain inferences equivalent to those from the actual posterior distribution $P(h(\theta) \mid y)$. In reality we can generate only $S^{(m)}$, so the best inferences attainable will be those based on the reduced information in the posterior $P(h(\theta) \mid S^{(m)})$. Perhaps we should focus our efforts on approximating $P(h(\theta) \mid S^{(m)})$.

Rubin's (1987) rules for combining point and variance estimates from a multiply-imputed dataset are based on this type of argument. Multiple imputation (MI) assumes that we have m *independent* draws of the missing data from their posterior predictive distribution given the observed data. The MI point estimate is simply a Rao-Blackwellized estimate of the posterior mean of $h(\theta)$, and the MI interval is a credible set based on an approximation to $P(h(\theta) \mid S^{(m)})$ where m may be very small. Allowances for the smallness of m are thus a built-in feature of the MI interval. Further discussion on the relationship between MI and Rao-Blackwellization is given by Schafer (1996). It may be profitable to consider how to approximate $P(h(\theta) \mid S^{(m)})$ for larger values of

m, where $S^{(m)}$ represents possibly *dependent* draws of some type of sufficient statistic arising from MCMC.

ROBERT L. STRAWDERMAN *(University of Michigan, USA)*

It is a pleasure to be asked to participate in this discussion of Professor Casella's paper, which does an excellent job in describing the interplay between Monte Carlo (MC) algorithms and statistical inference. MC itself is an inherently frequentist idea, with "long-run average" convergence properties being the primary justification behind its use in most applications. I find it particularly interesting that the vast majority of applications in which MC methods (particularly of the Markov chain variety, or MCMC) have been put to use is in solving Bayesian problems. Evidently, frequentist and Bayesian techniques complement each other more than is often explicitly recognized.

A prominent underlying theme of this paper is that MC methods are a very useful yet imperfect tool for statistical inference. Since MC methods have by definition a probabilistic basis, they can often be improved through clever statistical thinking. "Rao-Blackwellization" is indeed a clever method for optimizing an accept-reject (AR) algorithm; however, it is easy to see this procedure becomes impractical very quickly. Termwise conditional expectation is shown to be quite useful, particularly in conjunction with rescaling. The estimator δ_{Tr} (Eqn. 18) is really an importance sampler in disguise; its rescaled pure importance sampling competitor δ_{ISr} (Eqn. 20) is obviously so. It is known (e.g., Hesterberg, 1991, 1993) that simply dividing by the sum of the weights, while often effective, isn't necessarily an optimal procedure for improving importance-based sampling estimates. I wish to comment briefly on this aspect in somewhat more detail, with the particular objective of improving upon both δ_{Tr} and δ_{ISr} through the use of *control variates*. Then, I'd like to propose one possible solution to the problem that Professor Casella poses in Section 5.1.

Let $Y|N = n$ be a random variable having density $m(y)$ (Eqn. 15). Then, we may write $\tau = E_f[h(X)] = E_N[E_{Y|N}[h(Y)f(Y)/m(Y)]]$ by the usual importance sampling identity. Notice the similarity here to the weights used in calculating δ_{Tr}, hence the importance sampling

interpretation of δ_{Tr}. Setting $d(Y) = h(Y)f(Y)/m(Y)$, then obviously

$$E_N[E_{Y|N}[h(Y)f(Y)/m(Y)]] = \beta E_N[E_{Y|N}[c(Y)]] + E_N[E_{Y|N}[d(Y) - \beta c(Y)]]$$

for any function $c(Y)$ and some constant β. This is the key identity behind control variates in disguise; the optimal choice for β in terms of achieving minimum variance is $\beta = \text{cov}(d(Y), c(Y))/\text{var}(c(Y))$ (cf. Hesterberg, 1991). Ideally, the more correlated $c(Y)$ and $h(Y)f(Y)/m(Y)$, the larger the reduction in variance. This may be a difficult choice in practice; thus, for convenience, consider setting $c(Y) = d(Y)f(Y)/g(Y) = h(Y)f^2(Y)/(m(Y)g(Y))$; then, it is easy to see that $\mu_c = E[c(Y)] = E_g[h(Z)f^2(Z)/g^2(Z)]$, where Z has density $g(\cdot)$.

Now, let $\hat{\beta}$ be the slope of the regression of $d(y_i) = h(y_i)f(y_i)/m(y_i)$ on $c(y_i)$, $i = 1 \ldots n-1$ where $(y_1, \ldots, y_{n-1})$ are the first n-1 accepted and rejected rv's. Although y_i and y_j are correlated, each is an observation having marginal density $m(\cdot)$. I propose

$$\delta_{CV} = \hat{\beta}\mu_c + (\bar{d}_{n-1} - \hat{\beta}\,\bar{c}_{n-1})$$

as a competitor to δ_{Tr} and δ_{ISr}, Where $\bar{d}_{n-1}, \bar{c}_{n-1}$ respectively denote the sample averages. Note that if $y_i, i = 1 \ldots n-1$ were an iid sample, then δ_{CV} asymptotically achieves the minimum variance among linear estimators of the form $\beta\mu_C + (\bar{d}_{n-1} - \beta\bar{c}_{n-1})$. In practice, we may replace μ_c by an initial MC estimate $\hat{\mu}_c$, the latter usually being very quick to obtain since $g(\cdot)$ (the AR density) is generally easy to sample from. I reran a small portion of the simulation study done by Professor Casella (with code written in S-Plus) to investigate whether this new estimator provides any additional improvement. The results, represented as a percentage decrease in MSE over δ_{AR}, are summarized in Table 1.

The gains provided by δ_{CV} are impressive here, and have been essentially obtained via linear regression; there are few techniques which are more statistical than that! An interesting question here is the asymptotic relative efficiency of this procedure compared to full Rao-Blackwellization.

Turning now to the question posed in Section 5.1, we wish to determine a^* such that

$$\frac{1}{m}\int_{-\infty}^{a^*}\sum_{i=1}^{m}\pi(\theta|\boldsymbol{y}, \lambda_i)d\theta = \alpha$$

Table 1. *Estimating $E[h(X)]$ for X a Gamma random variable ($E[X] = 1/2$, 2500 simulated datasets via AR algorithm)*

AR Sample Size	$h(x)$	*Acceptance Rate 0.9* % Dec. in MSE δ_{Tr}	% Dec. in MSE δ_{ISr}	% Dec. in MSE δ_{CV}	*Acceptance Rate 0.3* % Dec. in MSE δ_{Tr}	% Dec. in MSE δ_{ISr}	% Dec. in MSE δ_{CV}
10	x	16.3%	19.2%	93.1%	63.2%	63.3%	99.6%
25	x	19.3%	21.0%	94.8%	68.7%	68.7%	99.7%
10	$x^{0.1}$	16.9%	19.8%	55.2%	62.1%	62.2%	93.3%
25	$x^{0.1}$	26.3%	26.6%	75.2%	68.2%	68.2%	94.3%

based on the Gibbs sequence $(\theta_1, \lambda_1), (\theta_2, \lambda_2), \ldots$. This problem can be immediately generalized to finding a^* such that $\int_{-\infty}^{a^*} g_m(\theta) d\theta = \alpha$, where

$$g_m(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{\phi(\theta_i|\lambda_i)}{f(\theta_i, \lambda_i)} f(\theta, \lambda_i)$$

for any proper conditional density $\phi(\theta|\lambda_i)$ having the same support as $\pi(\theta|\boldsymbol{y}, \lambda)$ and $f(\theta, \lambda) \propto \pi(\theta, \lambda)$, the latter being the joint posterior density of (θ, λ) given $\boldsymbol{y}$. The function $g_m(\theta)$ is the *importance weighted marginal density* (IWMD) estimator of Chen (1994), and reduces to $m^{-1} \sum_{i=1}^{m} \pi(\theta|\boldsymbol{y}, \lambda_i)$ for $\phi(\theta|\lambda) = \pi(\theta|\lambda)$. The ensuing proposal therefore covers both possibilities. The density estimate $g_m(\theta)$ may not integrate to 1 (cf. Chen, 1994, §5); it is useful to note here that the following will only require $g_m(\theta)$ to integrate to c for some $c > 0$, and thus no numerical renormalization of $g_m(\theta)$ is necessary.

Given a Gibbs sequence $(\theta_1, \lambda_1), (\theta_2, \lambda_2), \ldots, (\theta_m, \lambda_m)$, we can easily calculate the corresponding IWMD estimate. Suppose that m is reasonably large and that $\pi(\theta|\boldsymbol{y}) \approx c^{-1} g_m(\theta)$ is unimodal with $\hat{\theta} =$

$\text{argmax}_\theta g_m(\theta)$. Then, under some regularity conditions,

$$P\{\theta > a\} \approx \Phi(R_a) + \left\{ R_a^{-1} - \frac{(-k^{(2)}(\hat{\theta}))^{1/2}}{k^{(1)}(a)} \right\} \phi(R_a)$$

for $k(\theta) = \log g_m(\theta)$, $k^{(j)}(\theta) = \frac{d}{d\theta} k(\theta)$, and $R_a = \text{sign}(\hat{\theta})[2(k(\hat{\theta}) - k(a))]^{1/2}$ (cf. DiCiccio and Martin, 1993, Eqn. 5). An exactly analogous result obtains in any higher dimensional problem; that is, the formula is exactly the same in the case where a marginal probability calculation is desired for a single component of a vector-valued parameter.

Let $H(a;\alpha) = P\{\theta > a\} - \alpha$; note that $H(a;\alpha)$ is monotone in a. Replacing $P\{\theta > a\}$ by the tail probability approximation above, the resulting approximation is monotone in a away from the posterior mean and the extreme tails. Hence, a bisection algorithm will quickly solve $H(a^*;\alpha) = 0$ for a^*; the advantage of bisection over, say, Newton's method is that the former works without requiring derivatives. Use of this tail probability approximation requires maximization and taking derivatives of $k(\theta) = \log g_m(\theta)$. This should not be of great concern, and will typically not pose a problem in practice. For simplicity, suppose that we have calculated $\{(a_i, g_m(a_i)),\ i = 1 \ldots b\}$ on a reasonably fine grid $(a_1 \ldots a_b)$. Then, for example, to obtain an accurate estimate of $\hat{\theta}$ (the marginal posterior mode), one can fit a quadratic regression to $k(\theta)$ in a neighborhood about the approximate mode (i.e., , $\text{argmax}_{a_i}\ g_m(a_i)$), and then analytically calculate $\hat{\theta}_q$ (and also approximate $k(\hat{\theta}_q)$ and $k^{(2)}(\hat{\theta}_q)$) using the estimated regression equation (cf. DiCiccio *et al.*, 1996). Alternatively, we can take $\hat{\theta} = \text{argmax}_{a_i} g_m(a_i)$ and calculate all derivatives numerically. Each keeps in the spirit of constructing the answer only from the Gibbs sequence.

To illustrate this technique we reanalyzed data from Farewell and Sprott (1988). A mixture model was proposed for analyzing count data; the two-parameter (conditional) likelihood function is given there, as are asymptotic confidence intervals based on the MLE's of the model parameters. This particular example can also be found in Spiegelhalter *et al.* (1996, *BUGS* Examples Manual, Volume II, pp. 11-12), where Gibbs sampling is used to construct 95% posterior intervals for the model parameters, both of which are probabilities (p and θ, say) and are assumed independent. The intervals there are found by generating a Gibbs chain

based on 11,000 iterations (the first 1000 of which are treated as "burn-in"), and then marginal posterior intervals are respectively calculated via the empirical cdf's of the 10,000 iterates of p and θ.

The full conditionals are not "nice" in this problem, and it is advantageous to use the IWMD estimator. Based on the Gibbs output, I estimated the marginal densities of p and θ as discussed above; $\phi(\cdot|\cdot)$ was taken to be a Beta density with mean and variance matching the empirical mean and variance of the parameter whose marginal density was being computed. To calculate the posterior marginal HPD region for θ, I generated the IWMD estimate for θ on an equally-spaced grid of points (mesh = 0.01). Tail probabilities at any given point (away from the very extreme tail) were then calculated using the tail probability formula above. This was accomplished by setting $\hat{\theta} = \text{argmax}_{a_i} g_m(a_i)$ and then computing $k(\hat{\theta})$ and $k^{(j)}(\hat{\theta}), j = 1, 2$, the latter via standard formulas for numerical derivatives. Recalling that $H(a;\alpha) = P\{\theta > a\} - \alpha$, the equations defining the 95% marginal HPD limits are $H(\theta_U; 0.025) = 0$ and $H(\theta_L; 0.975) = 0$. As an approximation to θ_U, I used $\theta_{U,I} = 0.5(a_1 + a_2)$ where $a_1 = \text{argmax}_a\{H(a; 0.025) \geq 0\}$ and $a_2 = \text{argmin}_a\{H(a; 0.025) \leq 0\}$; θ_L was determined similarly. The results are summarized in Table 2.

Table 2. *Comparison of Highest 95% MPD Regions for PVC data from Farewell and Scott (1988) computed based on 10,000 Gibbs iterates*

Parameter	MLE	*BUGS*	Proposed method	Exact†
θ	(0.300, 0.810)	(0.289, 0.823)	(0.305,0.805)	(0.3012,0.8037)
p	(0.270, 0.520)	(0.264, 0.514)	(0.265,0.515)	(0.2693,0.5151)

† based on renormalized IWMD estimate using 32-point Gaussian quadrature

The DiCiccio and Martin formula performs extremely well here, given that it is based completely on numerical approximations. For comparison, the quadratic regression method (based on a symmetric window of 10 points containing $\text{argmax}_{a_i} g_m(a_i)$) mentioned earlier yields identical answers to the precision reported here.

REPLY TO THE DISCUSSION

First of all, I want to thank the organizers of the meeting, Professors José Bernardo and Elias Moreno for providing such a lively forum for the exchange of many stimulating ideas. Then I want to thank all of the discussants, who have raised so many interesting points and concerns that I could keep myself and my students busy for many years trying to answer them. For now, I will only try to provide a few thoughts. Since we are all working under time constraints, many of my comments will not be as complete as I would like them to be, but I still hope they will add something. (Indeed, I wish that I had more time to fully digest all of the extremely interesting points raised by the discussants, many with which I wholeheartedly agree.)

It seems to be most logical to arrange my responses by subject rather than people, and I will start with the one that, perhaps evoked the most comments.

1. *The Bayes/Frequentist Synthesis*

It is gratifying that most people agree that, as statisticians, our main concern should be to solve problems as best as we can, and use whatever tools are available. Such are the sentiments of Professors Berger, Gustafson and Wasserman, Ferrándiz, Peña, and Strawderman, with Berger raising a particularly interesting point. My Examples 1 and 2 indeed show how the *tools* of one approach can help the other approach. The question of the inference, to me, is a somewhat different one in that the appropriate inference is a decision of the experimenter. Although I believe that, in many cases, the frequentist inference is the appropriate one, there are situations where a Bayesian inference is more appropriate. Again, even in the question of inference, there is no (or, at least, little) need to argue. In consultation with the statistician, the experimenter should decide on the appropriate inference, and the statistician should help the experimenter make that inference in the best way possible.

The point is that we shouldn't have Bayesian and frequentist statisticians, we should have Bayesian and frequentist inference, to be appropriately used and recommended by all statisticians.

2. *Computational Algorithms*

At the very least, I am heartened that some of this work has resulted in people being sensitized (but not in the sense of Professor Meng) to the impact of the algorithm on the inference. The concerns of Professor Peña are well founded, and the guidelines of Professor Rios Insua are quite important. As Professor Schafer points out, focusing on the algorithm may be one step removed from our ultimate purpose, but it is an important step. As we will see in Section 4.2, problems can appear even with seemingly reasonable MC estimators. But even more importantly, I believe that we are all beginning to approach theoretical problems in a new way, always thinking of the computations, and being concerned more with algorithms than theorems. Such an approach can only enhance our thinking and broaden our influence.

3. *Posterior Distributions*

The power variance priors of model (4) are mainly chosen because (i) experimenters tend to believe that improper priors reflect impartiality and (ii) they result in easy to simulate conditionals. As Professor Peña notes, the Jeffreys priors considered by Ibrahim and Laud (1991) indeed give proper posterior distributions, as will Professor Bernardo's reference priors, as they both control the tail at zero. Any reanalysis with these priors will result in coherent inferences, the only drawback being that the conditional distributions are not as easy to sample from. However, the inferences are definitely superior.

The popularity of the power prior is an example of the algorithm overshadowing the statistics. Experimenters were so keen to make the Gibbs sampler work that they forgot to check the fundamentals of the model. Moreover, choosing $a = b = 0$ in (4), which usually is justified through an invariance argument, is extremely unfortunate as, for example, $a = b = 1/2$ would yield easily obtained conditionals and proper posterior distributions.

Many discussants had extremely interesting comments and concerns about this topic. I can loosely group those concerns in the following subsections.

3.1. Incompatibility. The property of *compatibility* of densities has received a lot of comment, and I am heartened that the discussants feel that this property is as important as Jim Hobert and I do. I should first

mention that, in response to Professors García-López and González, the results of Theorem 2 hold for the Data Augmentation Algorithm, which can be considered bivariate (but possibly vector valued) Gibbs sampling.

Professor Meng's discovery of his equation (1) is very interesting. It is one of those neat facts that, in hindsight, are totally obvious but, in foresight, are maddeningly difficult to see. I am not aware of the history of the representation, but had seen it presented as a special case of the Hammersley-Clifford Theorem by Robert (1996, Section 5.1.4, Lemma 5.3). It is a wonderful learning equation.

Professor Liu's comments on incompatible densities are also very interesting, and I would like to discuss how they fit in with Theorem 2. In Liu's notation, f_1 and f_2 are proper densities which are not functionally compatible, but $T_x(x, x') = \int f_1(x|y)f_2(y|x')dy$ and its counterpart T_y define positive recurrent transition functions. In some sense this is "almost as good" as being compatible, as there will exist limiting probability distributions. Thus, although the inference is more complicated, there is a legitimate inference to be recovered here.

The key fact that gets these limiting distributions is that T_x and T_y define positive recurrent Markov chains. But what happens in the functionally compatible (but *not* compatible) case? In this case, again using Liu's notation, the marginal distributions π_1 and π_2 will not be proper. This follows because, for example, $\int \pi_1(y)dy = \int\int \pi_1(x, y)dxdy$ and, by Theorem 2, this latter integral must be ∞, or else the densities would be compatible. Thus, the situation illustrated by Professor Liu cannot occur in the functionally compatible, but not compatible, case. As an example, consider the exponential densities of Example 3, which are not compatible. There we have

$$T_x(x, x') = \int ye^{-xy}x'e^{-yx'}dy = \frac{x'}{(x + x')^2},$$

and the invariant distribution is $\pi_1(x) = 1/x$, which is easily verified to be the solution to $\pi_1(x) = \int T_x(x, x')\pi_1(x')dx'$, and is not a proper distribution.

Perhaps Professor Liu has uncovered a property more fundamental than compatibility. Compatibility will insure the existence of one limiting probability distribution, but if T_x and T_y define positive recurrent Markov chains there will be a collection of limiting probability distributions. In some cases, this may be enough to recover a reasonable

statistical inference. Which leads us to subchains and submodels and the discussions of Professors George and Berger.

3.2. Inferences from an Improper Posterior. The arguments of Professor George are not compelling, because in every case the full Gibbs chain clearly contains extraneous pieces. To put it more formally, suppose that we are interested in inference about the parameter β, and have a model that results in the full, improper posterior $\pi(\alpha, \beta|y)$, where α is another parameter of the model, considered as a nuisance parameter when the inference is about β. Inferences about β would be based on the marginal posterior $\pi(\beta|y)$, which should satisfy

$$\pi(\beta|y) = \int \pi(\alpha, \beta|y) d\alpha.$$

If so, then it is impossible for $\pi(\beta|y)$ to be proper, as

$$\int \pi(\beta|y) d\beta = \int \pi(\alpha, \beta|y) d\alpha d\beta = \infty.$$

Thus there is no meaningful inference about the parameter β that can be recovered from the full model. (I also suspect that any inference about β in this model would be *incoherent* in the sense of Heath and Sudderth 1989).

So what about the experience of Berger, and the examples of George? These are instances in which there is reason to abandon the full model. That is, the transformations of George, and the "identifiability" of Berger are procedures for changing the model. In my illustration above, the parameter α would be somehow eliminated, and only β would be considered, with a proper $\pi(\beta|y)$. So my point is that if a model results in an improper full posterior, there is no lower dimensional inference based on the *full model* that can make sense. However, there may be a lower dimensional model that makes sense. I have no problem with this solution, but realize that the model is being changed in a fundamental way; we are not recovering anything from the improper posterior distribution. The interesting procedure discussed by Meng, that of *recursive deconditioning* seems to be an excellent candidate for searching for such lower dimensional models

3.3. Fixing Impropriety. If the posterior distribution is improper, an obvious fix is to replace it with a sufficiently "vague" proper prior that

is close to it. This is the spirit of Berger's suggestion to constrain $\sigma > 0$ in Example 4. As the values of σ do not spend too much time near the singularity at zero (as noted at the end of Example 4), the constrained prior might be a reasonable approximation here. However, such a fix may not always work. Natarajan and McCulloch (1996) investigate the effects of replacing improper priors with vague, proper priors and find that there is no happy medium between "proper but diffuse" and "improper". In particular, in situations where the posterior does not exist, the Gibbs sampler can break down before the prior becomes diffuse enough to yield estimates that are reasonable approximations to the MLE. But I guess that my sentiments on this problem are most in line with Gustafson and Wasserman, when they state that to use a proper vague prior is "..simply to approximate an ill defined solution".

The behavior of this Gibbs chain also answers the comment of Rios Insua, who expected more mass near zero. Such behavior was not exhibited by the chain, even with many restarts and many long runs (which should have eliminated any problems due to sample size or starting points – a concern of García-López and Gonzalez). This also illustrates, once again, the (apparent) futility of trying to have the Gibbs output check itself for propriety.

4. *Rao-Blackwellization*

The technique of Rao-Blackwellization has expanded beyond the original idea of conditioning on a sufficient statistic. Indeed, in my thinking, it has expanded to encompass a class of techniques that aim at improving estimators by taking advantage of the structure of the problem in whatever manner is available.

I don't believe that we have returned to the status quo, as stated by Berger. Even in situations where we end up with the same procedures, we also end up learning a lot (the gains of Rao-Blackwellization can be huge, and easy to obtain) and have not always returned to the status quo (the full Rao-Blackwellized estimator is still the only one to achieve substantial gains while retaining unbiasedness.) Although Ferrándiz rightly points out that the Rao-Blackwellization in the paper only applies to algorithms with ancillary random variables, the general approach goes far beyond this case. Perhaps the most important contribution is that we have stimulated thinking to search for better ways to process the output,

searches that have resulted in procedures such as those put forth by Professors Phillipe and Strawderman which, in our expanded definition, are again some sort of Rao-Blackwellization.

Rao-Blackwellization is a type of smoothing, and the advantages of such smoothing are well documented. I was particularly interested in the interpretations of Professor Dawid that cast new light on importance sampling, accept-reject, and weighted averages. Dawid's discussion clearly shows the drawback of the naive accept-reject average, and the advantage of the "Rao-Blackwellization" brought on by importance sampling.

Before replying to some of the other comments on Rao-Blackwellization, I would like to elaborate on a small point that has intrigued me for a while. Although it is clear that importance sampling is a desirable technique when compared to accept-reject or Metropolis-Hastings averages, its usefulness in the Gibbs sampler is not at all clear. For a bivariate Gibbs sampler (X_1, Y_1), (X_2, Y_2), $\cdots$, (X_m, Y_m), where we generate $X_i \sim f(x|Y_i)$ and $Y_{i+1} \sim f(y|X_i)$, a Gibbs estimate $\delta_G = \frac{1}{m}\sum_{i=1}^{m} h(X_i)$ has an importance sampling counterpart

$$\delta_{IS} = \frac{1}{m}\sum_{i=1}^{m} \frac{f(X_i)}{f(X_i|Y_i)} h(X_i)$$

(ignoring the possibility that the marginal $f(x)$ may not be computable). An interesting fact is that

$$E\left[\frac{f(X_i)}{f(X_i|Y_i)} h(X_i) \,\middle|\, X_i\right] = h(X_i),$$

so, here, the naive Gibbs average is the "Rao-Blackwellization" of the importance sampling estimate. However, dominance does not follow immediately, as there are covariances to contend with. But, I can show that for $m = 2$, $\text{var}(\delta_G) < \text{var}(\delta_{IS})$. Thus, this may be saying that the Gibbs sampler is already "smooth enough", and there is no room for further smoothing.

4.1. Termwise Rao-Blackwellization. First a short comment on the discussions of Liu and Dawid about termwise conditioning, and the importance of the stopping rule–it cannot be ignored. The stopping rule brings us the fact that the accept-reject estimator (10) is both unbiased and

"correct for constants". This is perhaps more clear when the estimator is written in the form (9), which can only be done with the knowledge of the value of t, that is, with knowledge of the stopping rule. The estimator δ_{IS} of Liu's discussion, that is,

$$\delta_{IS} = \frac{1}{n}\sum_{i=1}^{n} w(y_i)h(y_i) \qquad (R1)$$

cannot be directly related to either (9) or (10). It is a Rao-Blackwellization of

$$\delta_0 = \frac{1}{n}\sum_{i=1}^{n} I[U_i \leq w(y_i)]h(y_i)$$

under independent sampling and

$$\begin{aligned}
\text{var}(\delta_0) &= \text{var}[E(\delta_0|Y_1, \cdots, Y_n)] + E[\text{var}(\delta_0|Y_1, \cdots Y_n)] \\
&= \text{var}[\sum_{i=1}^{n} E(\delta_{0_i}|Y_i)] + E[\text{var}(\delta_0|Y_1, \cdots Y_n)] \\
&= \text{var}[\delta_{IS}] + E[\text{var}(\delta_0|Y_1, \cdots Y_n)] \\
&\geq \text{var}[\delta_{IS}].
\end{aligned}$$

But this does not prove dominance of (R1) over δ_{AR} of (10) and, indeed, this is not the case as δ_{AR} will dominate for constant functions as indicated by Table 2. So, in fact, without correcting for constants, or taking into a ccount the stopping rule, neither δ_{IS} nor δ_0 are particularly attractive estimators.

Professors Liu and Dawid also make similar points about the desirability of using weights based on *marginal* chains, where possible. The marginalization seems to smooth things out, and make it sometimes possible to achieve variance reduction. However, there are some unforeseen pitfalls here–a built in computational difficulty in the marginalization. There is a need for trade-off in that the original algorithms will often replace an analytic calculation with computer time and random variable generation, and the marginalization may require a difficult analytic calculation, a point noted by Liu. For example, the proposal of Dawid, which seems to carry along with it some excellent variance reduction potential, also carries along a large computational burden. The following simple example was pointed out by Christian Robert, where we take

$\pi(y) \propto \exp(-y^2/2)$, $q(y|x) \propto exp(-[x^2 + y^2]/2)$ and the resulting $\alpha(x, y) = \min\{\pi(y)q(x|y)/\pi(x)q(y|x), 1\}$, the usual Metropolis-Hastings choice. We then get a $\beta(x)$ of the form

$$\beta(x) = \Phi(|x| - x) - \Phi(-|x| - x) + \frac{\exp(x^2/4)}{\sqrt{2}} \left\{1 - \Phi[\sqrt{2}(|x| - x)] + \Phi[-\sqrt{2}(|x| + x)]\right\}$$

making for a difficult simulation algorithm. Perhaps this problem should be approached using decision theory, where we balance ease of computation with variance reduction through a loss function.

4.2. Subtleties. Next, I would like to elaborate on the point made by Gustafson and Wasserman about the failure of the average of conditional densities (ACD) to accurately estimate the marginal. At first, their example was bewildering to me, and there seemed to be no reason for such behavior. To better understand the "paradox" I reduced it to bare essentials, and learned the following. The failure of the ACD estimate has nothing to do with Gibbs sampling, impropriety, or Markov chains. It is, in fact, a failure to satisfy the assumptions of the Lebesgue Dominated Convergence Theorem!

Consider that in their example all of the relevant distributions are proper, and the Ergodic Theorem applies. Thus, if we obtain the random variables $u_1, u_2, \cdots$, we must have for each t

$$\frac{1}{m}\sum_{i=1}^{m} \pi_{\sigma^2|u,y}(t|u^{(i)}, y) \rightarrow \int \pi_{\sigma^2|u,y}(t|u, y)m(u|y)du, \qquad (R2)$$

where $m(u|y)$ is the proper marginal distribution of u. So (R2) holds for each t in the Gustafson/Wasserman example. It seems that there is a real mystery as to why the convergence fails at 0. But a little reflection brings an interesting realization. Write

$$\pi(0|y) = \lim_{t\rightarrow 0} \pi_{\sigma^2|y}(t|y) = \lim_{t\rightarrow 0} \int \pi_{\sigma^2|u,y}(t|u, y)m(u|y)du.$$

At $t = 0$, indeed for any $t = t_0$, the Monte Carlo sum converges to

$$\frac{1}{m}\sum_{i=1}^{m} \pi_{\sigma^2|u,y}(t_0|u^{(i)}, y) \rightarrow \int \pi_{\sigma^2|u,y}(t_0|u, y)m(u|y)du = \int \lim_{t\rightarrow t_0} \pi_{\sigma^2|u,y}(t|u, y)m(u|y)du.$$

Thus, when we construct a Monte Carlo sum such as in (R2), we are implicitly interchanging the order of limit and integration! It is straightforward to check that Dominated Convergence will hold here for every $t_0 > 0$, but fails at $t_0 = 0$. This example illustrates that things can go wrong even when all distributions are proper.

4.3. Other Estimates. Comparing the performance of Rao-Blackwellization to a weighted bootstrap, or double bootstrap, as suggested by García-López and González, would be an interesting endeavor. As these procedures are related to importance sampling, we would expect reasonable performance and perhaps easy implementation. I hope to look into this in the future.

There were other very interesting competitors to the Rao-Blackwell improvement suggested by other discussants. First, I would like to further explore the control-variate estimator proposed by Strawderman, and try to understand why it does so incredibly well. The simple answer seems to be that it is based on a much bigger sample size. But the more interesting answer is that it takes even better advantage of the algorithmic construction.

I think of control variates as finding the appropriate unbiased estimator of zero. To improve on an estimator $\delta_0(x)$ by the method of control variates, we find another estimator $u(x)$, with *known* mean μ, and construct $\delta_1(x) = \delta_0(x) + b[u(x) - \mu]$ for some constant b. Then δ_0 and δ_1 have the same expected value, and $\text{var}(\delta_1) = \text{var}(\delta_0) + \text{var}(u) + 2\text{cov}(\delta_0, u)$. If we choose b to have the optimal value $b = -\text{cov}(\delta_0, u)$, then we achieve the maximal variance reduction $\text{var}(\delta_1) = (1 - \rho^2)\text{var}(\delta_0)$, where ρ is the correlation between δ_0 and u. Strawderman has given us a methodology for implementing such a control variate scheme in any importance sampler. And why does it do so much better? The answer lies in his calculation of $\hat{\mu}_C$. In a control variate scheme, this is a known parameter, and Strawderman estimates it by taking a very large sample from g. So, in effect, his estimator is based on a much larger sample size than δ_{Tr} or δ_{ISr}. Is this an unfair comparison? You bet it is! Is this an unfair estimator. No! In fact, it shows us another clever way of recycling the rejected random variables! This control variate scheme deserves further investigation. I would be very interested in seeing how it compares to δ_{Tr} or δ_{ISr} when we keep the number of generated random variables the same for each estimator.

The discussion of Professor Phillipe is literally brimming with ingenious ideas that not only yield new (and seemingly excellent) estimators, but also illustrates the benefits of intertwining algorithmic and statistical thinking. Her Riemann sum estimator (1) appears to be a serious competitor to all of the other estimators developed in these pages, but I think the most interesting developments are in her subsequent estimator, where the instrumental density g is chosen to satisfy the boundedness requirements of her Propositions 1 and 2. What a terrific blending of algorithms and theory! The use of the Gibbs average as a substitute for the marginal also has nice potential, although one must be on guard for difficulties such as those illustrated in Section 4.2.

5. *Other Concerns*

5.1. Multiple Paths. The question of multiple path Gibbs sampling was raised by both Bernardo and García-López and González, although in different contexts. Firstly, the number of paths used in the Gibbs sampler will not have any impact on propriety or compatibility, as these are properties of the underlying model, and the manner in which we observe the model cannot have any bearing. The question of how multiple paths can affect the variance of our estimate is also an interesting one, and prompted me to write the following.

Suppose that we have data Y, and want to calculate an estimate $\delta(Y)$ of $\tau = E[\delta(Y)]$. Using a Monte Carlo algorithm to calculate $\delta(Y)$, we obtain an output string from the algorithm, a sample T of length k, and calulate $\delta_k(Y)$ as our approximation of $\delta(Y)$. Note that we could refer to $\delta(Y)$ as $\delta_\infty(Y)$, the value of the estimate based on an infinite sample from our algorithm, that is, a sample T_∞ of infinite length. We then also have that $E[\delta_k(Y)|T_\infty] = \delta(Y)$. Now suppose that we run the algorithm many times (for example, a multiple path Gibbs sampler), and let $T_1, \cdots, T_m$ be m independent output strings from the algorithm, each of size k. For each T_i calculate the values $\delta_k^{(i)}$ and take as our estimate $\bar{\delta}_k = \frac{1}{m}\sum_{i=1}^m \delta_k^{(i)}$. The following variance analysis, which may be similar in spirit to those discussed by Schafer, should apply whether we are considering Bayesian or frequentist measures.

The variance of $\bar{\delta}_k$ is given by

$$\begin{aligned} \text{var}[\bar{\delta}_k(Y)] &= \text{var}(E[\bar{\delta}_k(Y)|T_\infty]) + E[\text{var}(\bar{\delta}_k(Y)|T_\infty)] \\ &= \text{var}[\delta(Y)] + \frac{1}{m}E[\tau_k^2] \end{aligned} \tag{R3}$$

where $\tau_k^2 = \text{var}(\delta_1^{(i)}|T_i)$, the variance that is only due to the algorithm, and is not due to the model. Now we can see the effect of multiple paths (m) and increasing the length of the chain (k). As $k \to \infty$, $\tau_k^2 \to 0$, so increasing the length of the chain will reduce the variation due to the algorithm and also diminish the effect of Rao-Blackwellization (but, as we saw in Section 5.2, not erase it). However, increasing m, the number of paths, has no direct effect on τ_k^2, but still will reduce $\text{var}(\delta)$. But this latter situation is less desirable, as we should strive to eliminate the variation due solely to the algorithm (which is under our control). Thus, this naive analysis seems to show that there is less to be gained in variance reduction, whether the criterion is Bayesian or frequentist, from running multiple chains.

Equation (R3) may also answer the concern of Ríos-Insua that our stream of "endless data" eliminates the role of Bayesian statistics. Indeed, a more careful analysis of (R3), and the effects of changing k and m would almost certainly need some form of prior input to help balance the effects of the model and the algorithm.

5.2. Accurate Approximations. Professor Strawderman reminds me of one of my own lessons, that of not forgetting that we are statisticians with a large box of tools. He brings the methods of higher-order asymptotics to bear on the Gibbs sampler, showing that the DiCiccio/Martin tail probability approximation results in an extremely accurate approximation to the desired posterior probability in Section 5.1. Bravo. Professors DiCiccio and Wells also note the place for higher-order asymptotics, and make an interesting point about recovering a frequentist inference in the face of the Bayesian "catastrophe". Of course, whether the posterior distribution is proper has no bearing on the frequentist inference, which can always be made. However, under such catastrophic priors, such as $a = b = 1$, the Gibbs sampler cannot be used to produce reasonable frequentist inferences. Indeed, conjecturing based on the results of Nataran and McCulloch (1996), such catastrophic priors could leave us quite far from reasonable frequentist inference.

Also, as noted by DiCiccio and Wells, there is much interest now in "probability matching", or finding prior distributions (such as Welch-Peers) that result in posterior probabilities that match frequentist probabilities. Although such priors are necessarily improper, they also necessarily must result in proper posterior distributions, hence avoiding the impropriety problems. This suggests that probability matching could be a reasonable basis for choosing a default prior and should be acceptable to an experimenter as an "impartial" choice. Moreover, I think there is still room for Rao-Blackwellization for, at the very least, it will serve to minimize the error due solely to the Monte Carlo algorithm.

5.3. Decision Theory. It is quite gratifying that the mixing of Decision Theory with algorithmic performance is viewed favorably by many of the discussants. The sentiments of Ferrándiz perhaps most closely reflect my own, in that I am hopeful for many benefits from embedding the algorithm in the appropriate decision problem.

The research here is still in the beginning stages, so although we have interesting possibilities, there are still few definite recommendations. I have no answer for Berger on the performance of the optimal minimax scan, but it seems that the calculations of Professors DiCiccio and Wells hold promise that we are looking at a good criterion. They have provided more convincing evidence that the risk function does a more complete job in capturing the essentials of the Markov chain.

ADDITIONAL REFERENCES IN THE DISCUSSION

Andrews, R., Berger, J. and Smith, M. (1993). Bayesian estimation of fuel economy potential due to technology improvements. *Case Studies in Bayesian Statistics* (C. Gatsonis, *et al.* , eds.), 1–77. New York: Springer-Verlag.

Berger, J. O. and Bernardo, J. M. (1992). Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.). Berlin: Springer, 323–340.

Berger, J. and Strawderman, W. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.* **24**, 931–951.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.), Brookfield, VT: Edward Elgar, (1995), 229–263.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. B* **36**, 192–236 (with discussion).

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. A* **143**, 383–430.

Caracciolo, S., Pelissetto, A. and Sokal, A. D. (1990). Nonlocal Monte Carlo algorithms for self-avoiding walks with fixed endpoints. *J. Stat. Phys.* **60**, 7–53.

Chen, M. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89**, 818–824.

Christiansen, C. and Morris, C. (1995). Hierarchical Poisson regression modeling. *Tech. Rep.*, Department of Health Care Policy, Harvard.

Daniels, M. (1996). A prior for the variance in hierarchical models. *Tech. Rep.*, Department of Statistics, Carnegie Mellon University.

Daniels, M. and Gatsonis, C. (1996). Multilevel hierarchical generalized linear models in health services research. *Tech. Rep.*, Department of Health Care Policy, Harvard.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233 (with discussion).

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov Chains. *Ann. Appl. Probab.* **1**, 36–61.

DiCiccio, T. and Martin, M. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc. B* **55**, 305–316.

DiCiccio, T., Kass, R., Raftery, A. and Wasserman, L. (1996). Computing Bayes factors by combining simulation and asymptotic approximations. *Tech. Rep.*, Carnegie-Mellon University, Pittsburgh PA.

DuMouchel, W. (1994). Hierarchical Bayes linear models for meta-analysis. *Tech. Rep.* **27**, National Institute of Statistical Sciences.

Farewell, V. and Sprott, D. (1988). The use of a mixture model in the analysis of count data. *Biometrics* **44**, 1191–1194.

Ferrándiz, J., López, A., Llopis, A., Morales, M. and Tejerizo, M. L. (1995). Spatial interaction between neighouring counties: cancer data in Valencia, (Spain). *Biometrika* **51**, 665–678.

Gelfand, A. and Rubin, D. R. (1991). A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 625–631.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Trans. Pattern. Anal. Mach. Intelligence* **6**, 721–741.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. B* **54**, 657–699 (with discussion).

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90**, 909–920.

Green, E. J. and Strawderman, W. E. (1991). A James-Stein type estimator for combining unbiased and possibly biased estimators. *J. Amer. Statist. Assoc.* **6**, 416, 1001–1006.

Gidas, B. (1995). Metropolis-type Monte Carlo simulation algorithms and simulated annealing. In *Topics in Contemporary Probability and Its Applications*. (J. L. Snell, ed.). CRC Press.

Heath, D. and Sudderth, W. (1989). Coherent inference from improper priors and from finitely additive priors. *Ann. Statist.* **17**, 907–919.

Hesterberg, T. (1991). Weighted average importance sampling and defensive mixture distributions. *Tech. Rep.* **148**, Division of Biostatistics, Stanford University.

Hesterberg, T. (1993). Control variates and importance sampling for the bootstrap. *ASA Proc. the Statist. Computing Section,* ASA, Alexandria, VA, 40–48.

Hoaglin, D. and Andrews, D. (1975). The reporting of computation-based results in statistics. *The American Statistician* **29**, 122–126.

Justel, A. and Peña, D .(1996a). Gibbs sampling will fail in outlier problems with strong masking. *J. Comp. Graphical Stat.* **5**, 176–189.

Justel, A and Peña, D. (1996b). Bayesian unmasking in linear models. *Tech. Rep.*, Universidad Carlos III de Madrid.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing* **6**, 113–119.

Liu, J. S., Chen, R. and Wong, W. H. (1996). Rejection control and importance sampling. *Tech. Rep.*, Department of Statistics, Stanford University.

Liu, J., Wong, W. H. and Kong, A. (1992a). Correlation Structure and convergence rate of the Gibbs sampler: Applications to the comparison of estimators and augmentation schemes. *Tech. Rep.* **299**, University of Chicago.

Kemeny, J. G. and Snell, J. L. (1983). *Finite Markov Chains*. Berlin: Springer

MacEachern, S. N., Clyde, M. A., and Liu, J. S. (1996). Sequential importance sampling for nonparametric Bayes models: the next generation. *Tech. Rep.*, Department of Statistics, Stanford University.

Muller, P. and Ríos Insua, D. (1996). Issues in the Bayesian analysis of neural network models. *Tech. Rep., UPM*.

Natarajan, R., and McCulloch, C. E. (1996). Gibbs sampling with diffuse priors: a valid approach to data-driven inference? *Tech. Rep.* **BU-1313-M**, Cornell University. Under revision for *J. Comp. Graph. Statist.*

Ng, K. W. (1995). On the inversion of Bayes theorem. Talk presented to the *The 3rd ICSA Statistical Conference*, August 17-20, 1995, Beijing, China.

Peña, D. and Tiao G. C. (1992). Bayesian robustness functions for linear models. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press

Philippe, A. (1996). Processing simulation output by Riemann sums. *Tech. Rep.* **02**, Université de Rouen.

Peskun, P. H. (1973). Optimal Monte Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.

Ríos Insua, D., Ríos Insua, S. and Martin, J. (1997). *Simulation: Methods and Applications*. RA-MA. (In Spanish)

Robert, C. P. (1996). *Méthodes de Monte Carlo par Chaînes de Markov.* Paris: Economica.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method.* New York: Wiley.

Samaniego, J. F. and Renau, D. M. (1994). Towards a reconciliation of the Bayesian and frequentist approaches to point estimation. *J. Amer. Statist. Assoc.* **89**, 427, 947–957.

Schafer, J. L. (1996). *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall, (in press).

Smith, A. and Gelfand, A. (1992). Bayesian statistics without tears. *Amer. Stat.* **46**, 84–88.

Sokal, A. D. (1989). *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms.* Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50.* Cambridge: MRC Biostatistics Unit.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–880.

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42**, 385–388.

Tanner, M. A. and Wong, W. H. (1991). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 528–550, (with discussion).

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors by using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* **90**, 614–618.

Yakowitz, S., Krimmel, J. E. and Szidorovszky, F. (1978). Weighted Monte-Carlo integration. *SIAM J. Numer. Anal.* **15**, 1289–1300.