Rejoinder to the Discussion of "Statistical Inference and Monte Carlo Algorithms"

George Casella *
Cornell University

November 4, 1996

Abstract

The discussants have raised a number of interesting points, with particular attention to the Bayesian/frequentist synthesis, the use of Rao-Blackwellization, and the impact of improper posteriors. We respond to many of their concerns, and raise a few more.

*This research was supported by NSF Grant No. DMS-9625440. This is paper BU-1374M in the Biometrics Unit, Cornell University, Ithaca, NY 14853. Original paper presented in Grenada, Spain, September 27, 1996. File sp_discl.tex

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 Garcia-Lopez and Gonzalez, 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 in 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 Garcia-Lopez 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

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

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 $m = 2$, $\mathrm{var}(\delta_G) < \mathrm{var}(\delta_{IS})$. 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,

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

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

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

under independent sampling and

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

But this does not prove dominance of (1) 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 account 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 tradeoff 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

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

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

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

where $m(u|y)$ is the proper marginal distribution of u. So (2) 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

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

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

$$\frac{1}{m}\sum_{i=1}^{m} \pi_{\sigma^2|u,y}(0|u^{(i)}, y) \to \int \pi_{\sigma^2|u,y}(0|u, y)m(u|y)du = \int \lim_{t\to 0} \pi_{\sigma^2|u,y}(t|u, y)m(u|y)du.$$

Thus, when we construct a Monte Carlo sum such as in (2), 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 Garcia-Lopez and Gonzalez, 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 Garcia-Lopez and Gonzalez, 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_k 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 taks 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

$$\begin{aligned}
\operatorname{var}[\bar{\delta}_k(Y)] &= \operatorname{var}(E[\bar{\delta}_k(Y)|T_\infty]) + E[\operatorname{var}(\bar{\delta}_k(Y)|T_\infty)] \\
&= \operatorname{var}[\delta(Y)] + E[\operatorname{var}(\bar{\delta}_k|T_\infty)]) \qquad (3)\\
&= \frac{1}{m}\operatorname{var}(\delta_*) + \frac{1}{m}E[\tau_k^2],
\end{aligned}$$

where $\tau_k^2 = \operatorname{var}(\delta_0^{(i)}|T_i)$, 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 will reduce $\operatorname{var}(\delta_0)$ and $\operatorname{var}(\delta_*)$. 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 (3) may also answer the concern of Rios Insua that our stream of "endless data" eliminates the role of Bayesian statistics. Indeed, a more careful analysis of (3), and the effects of changing k and m would almost certain 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 can not 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.

6 Additional References

1. Natarajan, R., and McCulloch, C.E. (1996). Gibbs Sampling with Diffuse Priors: A Valid Approach to Data-Driven Inference? Biometrics Unit Technical Report BU-1313-M, Cornell University. Under revision for *J. Comp. Graph. Statist.*

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

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