Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

Discussion of paper by Mira and Baddeley

Richard L. Smith
University of North Carolina, U.S.A.
rls@email.unc.edu

1. INTRODUCTION

This very impressive paper brings together an enormous number of ideas from seemingly disconnected fields. There are frequentist ideas associated with estimating equations for general stochastic processes, and approaches to optimality associated with the names of Godambe and Heyde; connections with the method of maximum likelihood and the method of moments, also pseudo-likelihood, general Bayes estimators and a new concept called the pseudo-posterior; connections with admissibility theory; connections with ideas of ordering Markov chains in terms of their rates of convergence in different metrics; and there is even a passing reference to the Stein (or Stein-Chen) method of probability approximation, which is usually regarded as a means of obtaining explicit upper bounds in rate-of-convergence problems.

With so many ideas interacting, it is difficult to identify any one specific theme to concentrate on for a discussion. In my discussion, I would like to give my own perspective on what is going on. I hope that by so doing, I can persuade other to read the paper and form their own perspective of what are the most important themes.

2. CONNECTION WITH BADDELEY (2000).

It is important to understand Baddeley's paper to set the present work in context. Baddeley's formulation was a classical one: we assume a random variable X from a density $f(x \mid \theta)$ where x or X lies in some sample space $\mathcal{X}$, and θ lies in a parameter space Θ. However we are thinking in terms of general stochastic models for which an explicit representation of f may be hard to compute.

Baddeley assumes that for each θ it is possible to construct a Markov chain $\{X_n,\ n = 0, 1, 2, ...\}$ for which $f(\cdot \mid \theta)$ is the stationary distribution. Assume it has a generator $\mathcal{A}_\theta S(x) = \mathrm{E}\{S(X_{n+1}) - S(X_n) \mid X_n = x\}$. Then for $X \sim f(\cdot \mid \theta)$,

$$\mathrm{E}_\theta\{\mathcal{A}_\theta S(X)\} \quad = \quad 0$$

for any sensible function $S(X)$.

It follows that setting

$$\mathcal{A}_\theta S(X) \quad = \quad 0$$

Richard L. Smith is Mark L. Reed III Distinguished Professor of Statistics at the University of North Carolina, Chapel Hill, U.S.A.

defines an *unbiased estimating equation for* θ in considerable generality. This approach is likely to be particularly appealing in cases where the exact formula for $f(\cdot \mid \theta)$ is intractable but it is relatively easy to define a Markov chain for which f is the stationary distribution — precisely the situation from which most MCMC methods begin. Thus although the viewpoint in Baddeley's paper is purely frequentist, it already contains a number of elements that are familiar to Bayesians who use MCMC.

First Example. Suppose $X = \{X_i,\ i \in \mathcal{I}\}$ is a Markov random field with density of the form $f(x \mid \theta) \propto \exp\{\theta V(x)\}$ where the normalizing constant is intractable.

Besag (1975) proposed to estimate θ by maximizing

$$\prod_{i \in \mathcal{I}} f(x_i \mid x_{-i}, \theta) \tag{1}$$

where x_{-i} denotes the set of all elements in x excluding x_i. Equation (1) is called the *pseudolikelihood* and the resulting estimator is the MPLE.

Baddeley showed that this is the time-invariance estimating equation for this model when the embedding Markov chain is Gibbs sampling.

Thus in this case a well-known estimator (the MPLE) has been derived as a time-invariance estimating equation. This suggests that the time-invariance estimating principle may be used to derive alternative estimators for a Markov random field or for more complicated stochastic models where suitable estimators either do not exist at all, or have been previously derived only through *ad hoc* arguments.

Second Example. Consider a stationary point process on a compact subset of $\mathbb{R}^d$ that has a density $f(x \mid \theta)$ defined for all possible realizations x with respect to a unit Poisson process. For this process, Besag *et al.* (1982), Jensen and Møller (1991) defined the pseudolikelihood function by extending the definition for Markov random fields.

However for this process there is another estimator called the Takacs-Fiksel estimator which depends on an arbitrary function h defined on the sample space $\mathcal{X}$. So this raises the question of what relationship exists between the two estimators.

Baddeley answered the question as follows: It is possible to define a spatial birth-death process of which X is the stationary distribution. If f is an exponential family and $S(x) = V(x)$ (the canonical statistic) then the time-invariance estimator is MPLE. On the other hand if $S = h$ (apparently we don't need an exponential family assumption in this case) then the time-invariance estimator is the Takacs-Fiksel estimator.

I believe this discussion highlights some of the major motivations for the approach. In some problems there are estimators (such as Takacs-Fiksel) that don't appear to be motivated by any likelihood-based approach. Time-invariance estimating equations provide a unifying perspective that allows such estimators to be related to maximum likelihood and pseudo-likelihood. There are also instances where the time-invariance approach leads to new estimators that have not been studied previously. Baddeley provides several other examples and makes a start on discussing properties such as consistency and asymptotic optimality.

3. NEW APPROACHES TO ESTIMATION

Given that background, what's new in the present paper?

I believe one can highlight two essentially new ideas:

- Type II estimating equations: Choose an antisymmetric function $T(x, y) = -T(y, x)$ and define $(\mathcal{F}_\theta T)(x) = \mathrm{E}_\theta\{T(X_n, X_{n+1}) \mid X_n = x\}$.

  Then for a reversible Markov chain, $\mathrm{E}_\theta\{(\mathcal{F}_\theta T)(X)\} = 0$. Hence setting $(\mathcal{F}_\theta T)(X) = 0$ defines an unbiased estimating equation for θ.

  If $T(x, y) = S(x) - S(y)$, this reduces to the Type I estimating equation.

  This extension is especially useful in cases where the embedding Markov chain is Hastings-Metropolis, because the acceptance probability for Hastings-Metropolis is a complicated function of both the existing and proposed new states, that cannot be reduced to a simple difference of two functions. However, there are a number of instances when an appropriate estimating equation can be expressed as a Type II EE.

- The second extension is to a Bayesian EE approach, in which we embed the *joint* distribution of X and θ in a Markov chain on $(\mathcal{X}, \Theta)$ — many examples of this lead to standard Bayes estimators.

As several examples in the paper show, these ideas significantly extend Baddeley's original approach. In particular, the Bayesian extension allows many Bayes estimators to be expressed as time-invariance estimators.

4. CONNECTION WITH ADMISSIBILITY THEORY

Brown (1971) showed that it is possible to characterize admissibility or inadmissibility of estimators of a multivariate normal mean in terms of recurrence or transience of an associated diffusion process.

Johnstone (1984, 1986) developed a similar characterization for the estimation of Poisson means, where in his case the associated Markov chain was a birth and death process.

Although these are celebrated papers, they are also highly technical and difficult to understand in even an intuitive way. On the other hand, the approach of Eaton (1982, 1992, 1997) is a less powerful but far more straightforward theory for determining when the generalized Bayes estimator derived from an improper prior ν on Θ is almost-ν-admissible, a slightly weaker concept than traditional admissibility. A simplified version of Eaton's recipe is as follows:

(i) For $\theta,\ \eta \in \Theta$, define

$$r(\theta \mid \eta) \quad = \quad \int_{\mathcal{X}} q(\theta \mid x) p(x \mid \eta) dx$$

where $p(x \mid \eta)$ is the likelihood and $q(\theta \mid x)$ is the posterior density given $X = x$ when the prior is ν.

(ii) Think of $r(\theta \mid \eta)$ as the transition density of a Markov chain on Θ.

(iii) Under suitable regularity conditions, the recurrence of this Markov chain implies almost-ν-admissibility of the generalized Bayes estimator.

(Eaton also comments on the converse property, i.e. when is it true that transience of the Markov chain leads to inadmissibility of the estimator? Apparently

there are no general theorems in this direction, but it nevertheless seems to be assumed that the result is generally true.)

Hobert and Robert (1999) defined an alternative Markov chain through the transition kernel

$$\tilde{r}(y \mid x) \quad = \quad \int_{\Theta} p(y \mid \theta) q(\theta \mid x) d\theta.$$

Note that this defines a Markov chain on $\mathcal{X}$. If we couple the $\mathcal{X}$ and Θ updatings together, we also get a joint Markov chain defined on $\mathcal{X} \times \Theta$.

Hobert and Robert showed that all three Markov chains (including Eaton's) are positive recurrent, null recurrent or transient together. In particular, in some cases it is possible to prove recurrence of the Markov chain on $\mathcal{X}$-space when the corresponding result on Θ-space would not follow from any known stochastic process results — a major motivation and justification for their approach.

The present paper shows how these Markov chains can be used to re-derive a number of known (Bayesian and frequentist) estimators. But it's unclear to what extent it leads to really new estimators. The maximum pseudo-posterior estimator (Section 8.2) is one example that clearly exploits this idea of updating both the $\mathcal{X}$ and Θ spaces in succession, but my impression is that further study will be needed to decide whether this really is a good idea.

5. ASYMPTOTIC PROPERTIES OF ESTIMATORS

The approach of this paper offers potentially a large number of estimators for a given stochastic model. Whether the estimator is derived from the original estimating equation approach of Baddeley (2000), or from one of the more Bayes-oriented schemes of the present paper, it is still natural to use frequentist properties such as asymptotic variance as a means of discriminating among different point estimators. Sections 4 and 9 of the paper refer to attempts to relate asymptotic properties of the estimators to order properties of the generating Markov chains; I would like to make some comments about that and to propose a small extension to one of the results of Mira and Baddeley (2001).

If we consider the estimator $\tilde{\theta}$ defined by solving $(\mathcal{A}_\theta S)(X) = 0$, then the Godambe-Heyde formula leads to the approximation

$$\mathrm{Var}(\tilde{\theta}) \quad \approx \quad \mathrm{E}\left\{\nabla_\theta (\mathcal{A}_\theta S)(X)\right\}^{-T} \mathrm{Cov}\{(\mathcal{A}_\theta S)(X)\} \mathrm{E}\left\{\nabla_\theta (\mathcal{A}_\theta S)(X)\right\}^{-1}.$$

This has numerous alternative names, including the "information sandwich formula".

If Θ is one-dimensional, the formula reduces to

$$\mathrm{Var}(\tilde{\theta}) \quad \approx \quad \frac{\mathrm{Var}\{(\mathcal{A}_\theta S)(X)\}}{\left[\mathrm{E}\left\{\frac{\partial (\mathcal{A}_\theta S)(X)}{\partial \theta}\right\}\right]^2} \tag{2}$$

Suppose now that $\{Y_n\}$ and $\{Z_n\}$ are two reversible Markov chains with the same stationary distribution π_θ. With obvious notation, we also let $\mathcal{A}_{Y,\theta}$ and $\mathcal{A}_{Z,\theta}$ denote the generators indexed by θ, and $\tilde{\theta}_Y$, $\tilde{\theta}_Z$ the resulting time-invariance estimators. If Y_n dominates Z_n in *covariance ordering* (Mira 2001), one of the consequences is

$$\mathrm{E}_\theta\{S(X)(\mathcal{A}_{Y,\theta} S)(X)\} \quad \leq \quad \mathrm{E}_\theta\{S(X)(\mathcal{A}_{Z,\theta} S)(X)\} \ \leq \ 0 \tag{3}$$

for any $S \in L_0^2(\pi)$, the class of square integrable functions having zero mean with respect to π (Mira and Baddeley, 2001).

Consider the case $S(x) = \frac{\partial}{\partial\theta} \log f(x;\theta)$. Then

$$\mathrm{E}_\theta\{S(X)(\mathcal{A}_{Y,\theta}S)(X)\} \quad = \quad -\mathrm{E}_\theta\left\{\frac{\partial}{\partial\theta}(\mathcal{A}_{Y,\theta}S)(X)\right\} \tag{4}$$

To see (4), we merely need to note that for a statistic $T(x,\theta)$ which is uniformly differentiable in θ,

$$\frac{\partial}{\partial\theta}E_\theta\left\{T(X,\theta)\right\} \quad = \quad E_\theta\left\{T(X,\theta)\frac{\partial \log f(X;\theta)}{\partial\theta}\right\} + E_\theta\left\{\frac{\partial T(X,\theta)}{\partial\theta}\right\}. \tag{5}$$

However for $T(x,\theta) = \mathcal{A}_\theta(S)(x)$, the left hand side of (5) is 0, and we then deduce (4).

By combining (2), (3) and (4), we deduce the following:

Proposition: If

(i) $\{Y_n\}$ dominates $\{Z_n\}$ in covariance ordering, for each θ, and

(ii) $\mathrm{Var}\{(\mathcal{A}_{Y,\theta}S)(X)\} \le \mathrm{Var}\{(\mathcal{A}_{Z,\theta}S)(X)\}$,

then (under uniform differentiability conditions) $\tilde{\theta}_Y$ is more efficient than $\tilde{\theta}_Z$ in the Godambe-Heyde sense.

This result differs from that in Mira and Baddeley (2001) only in that they assumed an exponential family, and that assumption seems to me unnecessary. This distinction could be important for extending the results to other kinds of spatial processes (such as those that arise in geostatistics) where an exponential family assumption would be unduly restrictive.

Nevertheless, this is still a limited result, since as shown in several examples by Mira and Baddeley (2001), assumption (ii) cannot be dispensed with. It seems that quite a bit more work is needed to understand exactly how properties of the embedded Markov chain translate to those of the time-invariance estimator.

6. CONCLUSIONS

This is a very stimulating paper that brings together numerous Bayesian and frequentist concepts and provides a very general perspective on estimation of stochastic processes. I congratulate the authors, and look forward to seeing further developments of their work.

7. ADDITIONAL REFERENCES

Besag, J.E. (1975), Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Besag, J., Milne, R. and Zachary, S. (1982), Point process limits of lattice processes. *Journal of Applied Probability* **19**, 210–216.

Jensen, J.L. and Møller, J., Pseudolikelihood for exponential family models of spatial point processes. *Annals of Applied Probability* **1**, 445–461.

Acknowledgement. I thank Dr. Antonietta Mira for sending me the unpublished manuscript Mira and Baddeley (2001) and for answering a number of other queries.