Milestones in Computational Statistics lecture

Gareth Roberts

"Bayesian Computation via the Gibbs Sampler and Related Markov-Chain Monte-Carlo Methods" by AFM Smith and GO Roberts, *J. Roy. Statist. Soc., 1993, 55, 3-23*

October 2023

History

Bayesian modelling and inference before 1990s was largely confined to very simple models.

Bayesian computation was carried out mostly by numerical integration schemes.

Some key papers which pointed to the future:

- 1984 Geman and Geman Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattn Anal. Mach. Intell., 6, 721-741.
- 1987 Tanner and Wong The calculation of posterior distributions by data augmentation (with discussion). J. Am. Statist. Ass., 82, 528-550.
- 1990 Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. J. Am. Statist. Ass., 85, 398-409.

The event

Part of special multi-paper event organised by Royal Statistical Society.

RSS Discussion (Read) papers arguably the most prestigious and high profile in Statistics. Usually ≤ 30 pages together with a published discussion (typically $3-5$ pages).

Given the exciting new arrival of MCMC methods in Bayesian Statistics, it was decided to organise a meeting dedicated to MCMC to be held in early 1993.

So RSS invited 4 papers including 2 from Nottingham: one by Adrian Smith and one from me.

But they had a hard deadline of ≤ 50 pages, and each of the papers came in at 20 pages!

Delicate negotiations

In the end the two Nottingham papers were "merged" and delicate discussions over pages led to:

- 21 for Smith and Roberts
- 14 for paper by Julian Besag and Peter Green
- 15 for Gilks, Clayton, Spiegelhalter and McNeil

Delicate negotiations

In the end the two Nottingham papers were "merged" and delicate discussions over pages led to:

- 21 for Smith and Roberts
- 14 for paper by Julian Besag and Peter Green
- 15 for Gilks, Clayton, Spiegelhalter and McNeil

And in the end there was 50 pages of discussion!, 10 time more than what was expected!

Despite almost bankrupting RSS the meeting was a great success!

Personal story

I did a PhD in Probability at Warwick and my first academic job was as lecturer in Nottingham.

The time around 1990 was very exciting- the buzz in Nottingham was very exciting.

Adrian Smith had built up a large group in developing early MCMC methods and applications in very diverse areas.

But there was no probabilistic expertise in the group, so I saw a natural niche for my skills.

What was it all about?

Why was everybody so excited?

What was it all about? The emancipation of Bayesian modelling. It can be thought of as the Computational Statistics revolution.

Peter Clifford who proposed the first vote of thanks (the first invited discussion) said:

At this meeting we have had the opportunity to hear a large amount about an important new area in statistics. It may well be remembered as the 'afternoon of the 11 Bayesians'. Bayesianism has obviously come a long way. It used to be that you could tell a Bayesian by his tendency to hold meetings in isolated parts of Spain and his obsession with coherence, self-interrogation and other manifestations of paranoia. Things have changed, and there may be a general lesson here for statistics. Isolation is counter-productive.

Structure of S+R

- Introduction to the Gibbs sampler and Metropolis-Hastings algorithm.
- Implementational issues: Output analysis (Gelman and Rubin, Raftery and Lewis, Roberts)
- Optimising efficiency: variance reduction, efficient simulation of conditionals
- Some theory and discussion of the gap between theory and advice for practitioners
- 4. Modelling: emphasis on sample based approaches and flexibility in what to estimate. Using samplers to estimate posterior and predictive densities

Structure of S+R (cont.)

- 4.3 Sensitivity analysis and the flexibility of Monte Carlo inference
- 5. Gibbs sampling Emphasis on flexibility: constrained models, hierarchical models, GLMs, time series.
- 6. Missing data and data augmentation: censored data, use of mixtures.
- 7. MH Gene mapping and modelling using deformable templates
- Appendix with some basic theory.

The Gibbs sampler

$$x_1^1 \text{ from } \pi(x_1|x_{-1}^0)$$

$$x_2^1 \text{ from } \pi(x_2|x_1^1, x_3^0, \ldots, x_k^0)$$

$$\ldots$$

$$x_k^1 \text{ from } \pi(x_k|x_{-1}^1)$$

Kernel density:

$$K_G(x^t, x^{(t+1)} = \prod_{l=1}^{k} \pi(x_l^{(t+1)} \mid x_j^t, j > l,\ x_j^{(t+1)}, j < l)$$

The Metropolis-Hastings algorithm

Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953)

We iterate the following

Given x^n:

1. Propose x' from proposal density $q(x^n, \cdot)$.
2. Accept x' with probability

$$\alpha(x^n, x') = \min\left\{1, \frac{\pi(x')q(x', x^n)}{\pi(x^n)(x^n, x')}\right\}$$

3. If we accept then set $x^{n+1} = x'$
4. Otherwise set $x^{n+1} = x^n$.

The resulting Markov chain is in detailed balance for π:

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x)$$

What made these algorithms natural for Bayesian statistics?

Bayes's theorem:

$$\pi(\theta|y) \propto L(y|\theta)p(\theta)$$

Neither of these algorithms requires knowledge of the normalisation constant.

Both methods offer the prospect of black/box computation, eg $\mathbf{E}(g(\theta|y)$ estimated by

$$E_N = \frac{\sum_{t=1}^{N} g(\theta^n)}{N}$$

or by

$$E_{n,N} = \frac{\sum_{t=n+1}^{N} g(\theta^t)}{N-n}$$

What made these algorithms natural for Bayesian statistics?

Bayes's theorem:

$$\pi(\theta|y) \propto L(y|\theta)p(\theta)$$

Neither of these algorithms requires knowledge of the normalisation constant.

Both methods offer the prospect of black/box computation, eg $\mathbf{E}(g(\theta|y)$ estimated by

$$E_N = \frac{\sum_{t=1}^{N} g(\theta^n)}{N}$$

or by

$$E_{n,N} = \frac{\sum_{t=n+1}^{N} g(\theta^t)}{N-n}$$

(How to choose n???)

The Gibbs sampler and conditional conjugacy

Eg $Y_1 \dots Y_n \sim N(\mu, \tau^{-1})$

Then given independent priors: $\mu \sim N(\mu_0, \tau_0^{-1})$, $\tau \sim$Gamma(α_0, β_0), then

the posterior distribution of (μ, τ) is no longer two independent Gaussian and Gamma components. Ie break down of conjugacy. However ...

The Gibbs sampler and conditional conjugacy

Eg $Y_1 \dots Y_n \sim N(\mu, \tau^{-1})$

Then given independent priors: $\mu \sim N(\mu_0, \tau_0^{-1})$,
$\tau \sim$Gamma(α_0, β_0), then

the posterior distribution of (μ, τ) is no longer two independent Gaussian and Gamma components. Ie break down of conjugacy. However ...

$$\pi(\mu|\mathbf{Y}, \tau) \sim N(??, ??)$$

and

$$\pi(\tau|\mathbf{Y}, \mu) \sim \text{Gamma}(??, ??)$$

(FILL IN THE GAPS...)

The Besag and Green paper

The paper had an emphasis on application in spatial statistics although many of the ideas it introduced/discussed had much more general application.

- Emphasis on spatial methods
- Integrated autocorrelation time
- Antithetic variable methods
- Multimodality and auxiliary variables
- Auxiliary variable methods: SW
- Application in agricultural field experiments

The Gilks et al paper

The paper had an emphasis on medical applications

But methodologically important for the use of directed graphical models, and l ed to the development of BUGS software

The legacy of the meeting and S+R in particular

- Complete change in ambition for Bayesian modelling
- Big increase in the influence and of Bayesian modelling
- Missing data/data augmentation - widespread applications (Mixture models, Hidden Markov models, censored data, models involving latent processes ...
- Graphical models
- The emergence of some basic theory (eg starting with Roberts and Smith 1994, Stoch. proc. Appl.)
- How to optimise MCMC algorithms for particular target distributions (scaling, reparameterisation, auxilary variables, choosing scan strategies ...)
- Software: WINBUGS, JAGS, PyMC, STAN, NIMBLE, Tensorflow. (PPL) PLUS many more specialised MCMC packages

Legacy: theory and methodology

- Output analysis: little has changed
- Model choice/averaging via RJMCMC
- Stability theory, CLTs, soft convergence, advances in Markov chain theory
- Scaling analysis of MCMC, non-asymptotic bounds
- Bespoke MCMC for multi-modality, non-Euclidian spaces
- Gradient based algorithms massive (ULA, MALA, HMC, PDMPs)
- Adaptive MCMC

What was NOT anticipated ...

... but led on naturally

Many things! But includes ....

- Big data
- Distributed algorithms
- Intractable likelihood
- Algorithmic union with SMC and other algorithms
- (Bayesian) Machine Learning