Analysis of the Gibbs sampler for a model related to James–Stein estimators

JEFFREY S. ROSENTHAL
Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1

Received January 1995 and accepted June 1995

We analyse a hierarchical Bayes model which is related to the usual empirical Bayes formulation of James–Stein estimators. We consider running a Gibbs sampler on this model. Using previous results about convergence rates of Markov chains, we provide rigorous, numerical, reasonable bounds on the running time of the Gibbs sampler, for a suitable range of prior distributions. We apply these results to baseball data from Efron and Morris (1975). For a different range of prior distributions, we prove that the Gibbs sampler will fail to converge, and use this information to prove that in this case the associated posterior distribution is non-normalizable.

Keywords: Convergence rate, James–Stein estimator, Gibbs sampler, Markov chain Monte Carlo

1. Introduction

Markov chain Monte Carlo techniques, including the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), data augmentation (Tanner and Wong, 1987), and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) have become very popular in recent years as a way of generating a sample from a complicated probability distribution (such as the posterior distribution in a Bayesian inference problem). A fundamental issue regarding such techniques is their convergence properties, specifically whether or not the algorithm will converge to the correct distribution, and if so how quickly. In addition to the many general convergence results (e.g. Tierney, 1994) and convergence diagnostics (e.g. Roberts, 1992; Mykland et al., 1992) which have been developed, a number of papers have attempted to prove rigorous bounds on convergence rates for these algorithms (Jerrum and Sinclair, 1989; Amit and Grenander, 1991; Frieze et al., 1994; Meyn and Tweedie, 1994; Lund and Tweedie, 1993; Mengersen and Tweedie, 1993; Frigessi et al., 1993; Rosenthal, 1993, 1995a, b). However, most of the results are of a quite specific and limited nature, and the general question of convergence rates for these algorithms remains problematic and largely unsolved.

In this paper we investigate the convergence properties of the Gibbs sampler as applied to a particular hierarchical Bayes model. The model is related to James–Stein estimators (James and Stein, 1961; Efron and Morris, 1973, 1975; Morris, 1983). Briefly, James–Stein estimators may be defined as the mean of a certain empirical Bayes posterior distribution (as discussed in the next section). We consider the problem of using the Gibbs sampler as a way of sampling from a richer posterior distribution, as suggested by Jun Liu (personal communication). Such a technique would eliminate the need to estimate a certain parameter empirically and to provide a 'guess' at another one, and would give additional information about the distribution of the parameters involved.

We consider, in particular, the convergence properties of this Gibbs sampler. For a certain range of prior distributions, we establish (Section 3) rigorous, numerical, reasonable rates of convergence. The bounds are obtained using the methods of Rosenthal (1995b). We thus rigorously bound the running time for this Gibbs sampler to converge to the posterior distribution, within a specified accuracy (as measured by total variation distance). We provide a general formula for this bound, which is of reasonable size, in terms of the prior distribution and the data. This Gibbs sampler is perhaps the most complicated example to date for which reasonable quantitative convergence rates have been obtained. We apply our bounds to the numerical baseball data of Efron and Morris (1975) and Morris (1983), based on batting averages of baseball players, and show that...

0960-3174 © 1996 Chapman & Hall
To be specific, we use a flat prior for \( \pi \), and use a conjugate prior distribution involving putting priors on \( \mu \) and \( A \), thus defining a posterior distribution to be estimated. The Bayesian approach then estimates \((K - 2)/\pi(\theta, A, Y)\)
conditionally independent. Here \( \theta \) is taken to be an 'initial guess' at \( \theta \)'s, and where \( (1 + A^2)^{-1} \) is replaced by its (unbiased) estimate \((K - 2)/(\sum(Y_i - \mu)^2)\).

In this paper we follow the suggestions of Jun Liu (personal communication) to regard \( \mu \) and \( A \) as further parameters to be estimated. The Bayesian approach then involves putting priors on \( \mu \) and \( A \), thus defining a posterior distribution

\[
\pi(\cdot) = \mathcal{L}(A, \mu, \theta_1, \ldots, \theta_K | Y_1, \ldots, Y_K).
\]

To be specific, we use a flat prior for \( \mu \), and use a conjugate prior of the form \( \text{IG}(a, b) \) for \( A \) (where \( \text{IG}(a, b) \) is the inverse gamma distribution with density proportional to \( e^{-b/2} \).) We shall see that the chosen values of \( a \) and \( b \) can greatly affect the properties of \( \pi(\cdot) \).

The remainder of this paper is thus concerned with the problem of sampling from the distribution \( \pi(\cdot) \) defined above, with

\[
Y_i, \theta_i \sim N(\theta_i, V) \quad (1 \leq i \leq K)
\]
\[
\theta_i | \mu, A \sim N(\mu, A) \quad (1 \leq i \leq K)
\]
\[
\mu \sim \text{flat prior on } \mathbb{R}
\]
\[
A \sim \text{IG}(a, b).
\]

To accomplish this sampling, we use a Gibbs sampler on \((A, \mu, \theta_1, \ldots, \theta_K)\). After choosing initial values \( A^{(0)}, \mu^{(0)}, \theta_1^{(0)} \) from some initial distribution, we follow the suggestion of Jun Liu (personal communication) by letting the Gibbs sampler update these variables repeatedly (for iterations \( k = 1, 2, 3, \ldots \)) by the (easily computed) conditional distributions

\[
A^{(k)} \sim \mathcal{L}(A | \theta_1^{(k-1)}, Y_i)
\]
\[
= \text{IG} \left( a + K - 1/2, b + \frac{1}{2} \sum (\theta^{(k-1)} - \bar{\theta}^{(k-1)})^2 \right);
\]
\[
\mu^{(k)} \sim \mathcal{L}(\mu | A^{(k)}, \theta_1^{(k-1)}, Y_i) = N(\bar{\theta}^{(k-1)}, A^{(k)}/K);
\]
\[
\theta_i^{(k)} \sim \mathcal{L}(\theta_i | A^{(k)}, \mu = \mu^{(k)}, Y_i)
\]
\[
= N \left( \frac{\mu^{(k)} V + Y_i A^{(k)}}{V + A^{(k)}}, V + A^{(k)} \right);
\]

where \( \bar{\theta}^{(k)} = \frac{1}{k} \sum \theta_i^{(k)} \). (From the point of view of the Gibbs sampler, this corresponds to treating \((A, \mu)\) as a single variable, and jointly updating \( (A^{(k)}, \mu^{(k)}) \sim \mathcal{L}(A, \mu | \theta_1 = \theta_1^{(k-1)}, Y_i) \).)

This definition specifies the distribution of the random variables \( A^{(k)}, \mu^{(k)}, \theta_i^{(k)} \) for \( k = 1, 2, 3, \ldots \). The Gibbs sampler is said to converge (in total variation distance \( \| \cdot \| \)) to the distribution \( \pi(\cdot) \), if

\[
\lim_{k \to \infty} \| \mathcal{L}(x^{(k)}) - \pi(\cdot) \| = 0
\]

where we have written \( x^{(k)} \) as shorthand for \((A^{(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_K^{(k)})\). If the Gibbs sampler does converge, then quantitative bounds on \( \| \mathcal{L}(x^{(k)}) - \pi(\cdot) \| \) are important because they can determine how long the algorithm should be run (i.e. how large \( k \) needs to be) before \( x^{(k)} \) can be regarded approximately as a sample from \( \pi(\cdot) \).

We note that this Gibbs sampler is similar in form to the Gibbs sampler for variance components models, as suggested in Gelfand and Smith (1990), and partially analysed (though not numerically) in Rosenthal (1995a). It is thus a realistic example of an applied use of the Gibbs sampler.