Journal of Multivariate Analysis 169 (2019) 61–80

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Fast Monte Carlo Markov chains for Bayesian shrinkage models with random effects

Tavis Abrahamsen [a,*], James P. Hobert [b]

[a] Department of Statistical Science, Duke University, P.O. Box 90251 Durham, NC 27708-0251, USA
[b] Department of Statistics, University of Florida, P.O. Box 118545 Gainesville, FL 32611-8545, USA

ARTICLE INFO

Article history:
Received 28 November 2017
Available online 1 September 2018

AMS subject classifications:
62F15

Keywords:
Bayesian shrinkage prior
Geometric drift condition
Geometric ergodicity
High-dimensional inference
Large p/small n
Markov chain Monte Carlo

ABSTRACT

When performing Bayesian data analysis using a general linear mixed model, the resulting posterior density is almost always analytically intractable. However, if proper conditionally conjugate priors are used, there is a simple two-block Gibbs sampler that is geometrically ergodic in nearly all practical settings, including situations where $p > n$ (Abrahamsen and Hobert, 2017). Unfortunately, the (conditionally conjugate) multivariate Gaussian prior on β does not perform well in the high-dimensional setting where $p \gg n$. In this paper, we consider an alternative model in which the multivariate Gaussian prior is replaced by the normal-gamma shrinkage prior developed by Griffin and Brown (2010). This change leads to a much more complex posterior density, and we develop a simple MCMC algorithm for exploring it. This algorithm, which has both deterministic and random scan components, is easier to analyze than the more obvious three-step Gibbs sampler. Indeed, we prove that the new algorithm is geometrically ergodic in most practical settings.

1. Introduction

The general linear mixed model is one of the most frequently applied statistical models. It takes the form

$$Y = X\beta + \sum_{i=1}^{m} Z_i u_i + e,$$

where Y is an observable $n \times 1$ data vector, X and $Z_1, \ldots, Z_m$ are known matrices, β is an unknown $p \times 1$ vector of regression coefficients, $u_1, \ldots, u_m$ are independent random vectors whose elements represent the various levels of the random factors in the model, and $e \sim \mathcal{N}_n(0, \lambda_0^{-1} I)$. The random vectors e and $u = (u_1^\top, \ldots, u_m^\top)^\top$ are independent, and $u \sim \mathcal{N}_q(0, \Lambda^{-1})$, where for each $i \in \{1, \ldots, m\}$, u_i is $q_i \times 1$, $q = q_1 + \cdots + q_m$, and $\Lambda = \lambda_1 I_{q_1} \oplus \cdots \oplus \lambda_m I_{q_m}$. We further assume throughout that $n \geqslant 2$, and that $q_i \geqslant 2$ for each $i \in \{1, \ldots, m\}$. For a book-length treatment of this model and its many applications, see McCulloch et al. [16].

In the Bayesian setting, prior distributions are assigned to β and $\lambda = (\lambda_0, \ldots, \lambda_m)^\top$. Unfortunately, any non-trivial prior leads to an intractable posterior density. However, if β and λ are assigned conditionally conjugate priors, then a simple two-block Gibbs sampler can be used to explore the resulting posterior density. In particular, if we assign a multivariate

* Corresponding author.
E-mail addresses: tavis.abrahamsen@duke.edu (T. Abrahamsen), jhobert@stat.ufl.edu (J.P. Hobert).

https://doi.org/10.1016/j.jmva.2018.08.014

Gaussian prior to β, and independent gamma priors to the precision parameters, then, letting $\theta = (\beta^\top, u^\top)^\top$, it is easily shown that given observed data y, $\theta|\lambda, y$ is multivariate normal, and $\lambda|\theta, y$ is a product of independent gammas. Since u is unobservable, it is treated like a parameter. Convergence rate results for this block Gibbs sampler can be found in Abrahamsen and Hobert [2].

Now consider this Bayesian mixed model in the high-dimensional setting where $p \gg n$. This situation can arise, e.g., in genetics and neuroscience where variability between subjects is most appropriately handled with random effects; see, e.g., [6,22]. While the model described above could certainly be used in this setting, the multivariate Gaussian prior on β is really not suitable. Indeed, when $p \gg n$, it is often assumed that β is sparse, i.e., that many components of β are zero. Unfortunately, the multivariate Gaussian prior for β will shrink the estimated coefficients towards zero, but not enough to produce an (approximately) sparse estimate of β. Additionally, when the components of β have varying magnitudes, the estimates of the "large" components will be shrunk disproportionately compared to the estimates of the "small" components. Below we propose an alternative prior for β that is tailored to the high-dimensional setting.

The well-known Bayesian interpretation of the lasso (involving i.i.d. Laplace priors for the regression parameters) has led to a flurry of recent research concerning the development of prior distributions for regression parameters (in linear models without random effects) that yield posterior distributions with high posterior probability around sparse values of β. These prior distributions are called continuous shrinkage priors and the corresponding statistical models are referred to as Bayesian shrinkage models; see, e.g., [4,5,8,19,20]. One such Bayesian shrinkage model is the so-called normal-gamma model of Griffin and Brown [8], which is given by

$$Y|\beta, \tau, \lambda_0 \sim \mathcal{N}_n(X\beta, \lambda_0^{-1} I_n), \quad \beta|\tau, \lambda_0 \sim \mathcal{N}_p(0, \lambda_0^{-1} D_\tau),$$

where $\tau = (\tau_1, \ldots, \tau_p)^\top$ and D_τ is a diagonal matrix with the τ_js on the diagonal. The precision parameter, λ_0, and the components of τ are assumed to be a priori independent gamma random variables with $\lambda_0 \sim \mathcal{G}(a, b)$ and $\tau_j \overset{\text{i.i.d.}}{\sim} \mathcal{G}(c, d)$ for all $j \in \{1, \ldots, p\}$. When $c = 1$, this model becomes the Bayesian lasso model introduced by Park and Casella [19]. We note that Bhattacharya et al. [4,5] show that, in terms of frequentist optimality, the Bayesian lasso has sub-optimal prior concentration rates in that it does not place sufficient mass around sparse values of β. Alternatively, shrinkage priors that have singularities at zero and robust tails, such as in the normal-gamma model with $c < 1/2$, have been shown to perform well in empirical studies.

In this paper, we propose and analyze an MCMC algorithm for a new Bayesian general linear mixed model in which the standard multivariate normal prior on β is replaced with the continuous shrinkage prior from the normal-gamma model. Our high-dimensional Bayesian general linear mixed model is defined as follows

$$Y|\beta, u, \tau, \lambda \sim \mathcal{N}_n\left(X\beta + \sum_{i=1}^m Z_i u_i, \lambda_0^{-1} I_n\right), \quad \beta|u, \tau, \lambda \sim \mathcal{N}_p(0, \lambda_0^{-1} D_\tau), \quad u|\tau, \lambda \sim \mathcal{N}_q(0, \Lambda^{-1}), \tag{1}$$

where λ and τ are a priori independent with $\lambda_i \overset{\text{ind}}{\sim} \mathcal{G}(a_i, b_i)$, for $i \in \{0, \ldots, m\}$, and $\tau_j \overset{\text{i.i.d.}}{\sim} \mathcal{G}(c, d)$ for $j \in \{1, \ldots, p\}$. This model can be considered a Bayesian analog of the frequentist, high dimensional mixed model developed by Schelldorfer et al. [25]. (Of course, it can also be viewed as a mixed version of the normal-gamma shrinkage model.) A similar sparse Bayesian linear mixed model has been proposed by Zhou et al. [27] for polygenic modeling. They assume a "spike and slab" prior consisting of a mixture of a point mass at 0 and a normal distribution for the components of β. However, it is well-known that spike and slab priors lead to MCMC algorithms that have convergence problems, especially when p is large [5,20].

Our model in (1), can easily be adapted to other global–local shrinkage priors as defined in [20]. For instance, we could place a Horseshoe-type prior on β simply by assuming $\sqrt{\tau_j} \sim C_+(0, \eta)$, where $C_+(a, b)$ represents the half-Cauchy distribution with location and scale parameters a and b, respectively. However, this particular model does not admit a similar 3-block Gibbs sampler because the full conditional density of τ does not have a standard closed form representation. While this is not a problem from an inference standpoint, a similar drift function analysis is not possible if the normalizing constants of the full conditional densities are not known. Makalic and Schmidt [15] have developed a closed form Gibbs sampler for the Horseshoe and Horseshoe+ standard regression models using auxiliary variables. It may be possible to prove that these samplers are geometrically ergodic, which would be an interesting problem for future work.

Recall that $\theta = (\beta^\top, u^\top)^\top$, and let $\pi(\theta, \lambda, \tau|y)$ denote the posterior density associated with model (1). This density is highly intractable and Bayesian inference requires MCMC to explore its posterior distribution. As we show in Section 2, the full conditional densities $\pi_1(\theta|\lambda, \tau, y)$, $\pi_2(\lambda|\theta, \tau, y)$, and $\pi_3(\tau|\theta, \lambda, y)$ all have standard forms, which means that there is a simple three-block Gibbs sampler available. Unfortunately, we have been unable to establish a convergence rate for this Gibbs sampler (in either deterministic or random scan form). However, we have been able to prove that a related hybrid algorithm does converge at a geometric rate. The invariant density of our Markov chain is

$$\pi(\theta, \lambda|y) = \int_{\mathbb{R}_+^p} \pi(\theta, \lambda, \tau|y)\, d\tau,$$