Brazilian Journal of Probability and Statistics
2012, Vol. 26, No. 4, 358–371
DOI: 10.1214/11-BJPS144

Bayesian statistics with a smile: A resampling–sampling perspective

Hedibert F. Lopes[a], **Nicholas G. Polson**[a] **and Carlos M. Carvalho**[b]

[a]*The University of Chicago Booth School of Business*
[b]*McCombs School of Business, University of Texas at Austin*

Abstract. In this paper we develop a simulation-based approach to sequential inference in Bayesian statistics. Our resampling–sampling perspective provides draws from posterior distributions of interest by exploiting the sequential nature of Bayes theorem. Predictive inferences are a direct byproduct of our analysis as are marginal likelihoods for model assessment. We illustrate our approach in a hierarchical normal-means model and in a sequential version of Bayesian lasso. This approach provides a simple yet powerful framework for the construction of alternative posterior sampling strategies for a variety of commonly used models.

1 Introduction

Bayesian inference about a parameter θ requires calculating conditional posterior beliefs $p(\theta|y)$ given data y. We assume that the data are generated from a probability model with marginal distribution, $p(y)$, conditional likelihood $p(y|\theta)$ and initial parameter beliefs $p(\theta)$. In most relevant models, the computation of the posterior $p(\theta|y)$, marginal $p(y) = \int p(y|\theta)\,p(\theta)\,d\theta$ and predictive $p(y_{n+1}|y^n)$ cannot be carried out analytically and approximations are obtained via simulation schemes. In this paper, we provide a simulation-based approximation to both posteriors and marginal likelihoods. We will reverse the logic in the sequential version of Bayes rule to provide a resample-sampling alternative to standard sampling-resample approaches. In this sense, we take the "Bayesian statistics without tears" analogy of Smith and Gelfand (1992) one step further and propose a "Bayesian statistics with a smile" approach to sequential Bayesian inference.

Our new look at Bayes's theorem then delivers a sequential, online inference strategy that should be exploited in the construction of effective posterior simulation schemes. These strategies are intuitive, easy to implement and teach, and deliver more for less as direct approximations to marginal likelihoods are also available. In contrast with Markov chain Monte Carlo (MCMC), our approach is inherently parallel—an important feature as more multiprocessor computational power becomes available.

Key words and phrases. Hierarchical models, MCMC, Gibbs sampling, Bayesian lasso, ANOVA.
Received October 2010; accepted February 2011.

Sampling-resampling and Markov chain Monte Carlo methods for drawing posterior samples are now commonplace. For example, Rubin (1987) and Smith and Gelfand (1992) develop sampling importance resampling (SIR) approaches whilst Gelfand and Smith (1990) develop Gibbs sampling and MCMC methods. A couple of issues remain with implementation of these methods; first, standard SIR methods suffer from particle impoverishment and, second, MCMC require repeated implementations in sequential problems. Moreover, it can be slow or hard to diagnose convergence. This can occur even in plain vanilla hierarchical models as highlighted in our examples.

The remainder of the paper is organized as follows. Section 2 develops our simulation-based approach to sequential inference in Bayesian statistics. The section also discusses the choice of priors, Monte Carlo error assessment and the computation of marginal likelihoods for model assessment. We illustrate our approach in Section 3 with two canonical examples: normals-means and Bayesian lasso. Section 4 concludes.

2 Bayes with a smile

2.1 Resampling-sampling

Given a model $p(y|\theta)$ and a prior distribution $p(\theta)$, the posterior distribution is $p(\theta|y) = p(y|\theta)p(\theta)/p(y)$. We wish to be able to sequentially draw sample from this distribution. Suppose that we currently have draws from $\theta_n^{(i)} \sim p(\theta|y^n)$, that is, the particle set $\{\theta_n^{(i)}\}_{i\geq 1}$ are draws from the posterior distribution of θ conditional on $y^n = (y_1, y_2, \ldots, y_n)$, the data up to the nth observation. We then wish to propagate these draws to $\theta_{n+1}^{(i)} \sim p(\theta|y^{n+1})$. In this notation, in order to do this, we construct an augmented vector Z_n^θ which provides conditional sufficient information for parameter inference. Specifically, we assume that Z_n^θ can be constructed so that we can draw samples from the distribution $p(\theta|Z_n)$ at each n. The vector Z_n^θ can depend on hidden states, parameters and data. By construction, $p(\theta|Z_n^\theta)$ is either analytically tractable or straightforward to simulate from. Suppressing the dependence on θ for the moment, given Z_n and samples $Z_n^{(i)} \sim p(Z_n|y^n)$, Rao-Blackwellisation will then determine

$$p(\theta|y^n) = E_{Z_n|y^n}(p(\theta|Z_n)) \quad \text{and} \quad p(\theta|y^n) \approx \frac{1}{N}\sum_{i=1}^{N} p\big(\theta|Z_n^{(i)}\big)$$

as our posterior approximation. See Gelfand and Smith (1990) for a further discussion of Rao-Blackwellisation and its efficiency gains.

We are still left with the problem of how to draw samples $Z_n^{(i)} \sim p(Z_n|y^n)$. Essentially, we have reduced the sequential learning problem of calculating $p(\theta|y^n)$ to a filtering problem of finding samples from the set of distributions $p(Z_n|y^n)$.

Panel A *Bayesian learning by resampling–sampling*

1. (*Resample*): Generate an index $k(i) \sim \text{Multi}(w_n^{(i)})$, where
$$w_n^{(i)} = \frac{p(y_{n+1}|Z_n^{(i)})}{\sum_{j=1}^N p(y_{n+1}|Z_n^{(j)})}.$$
2. (*Sample*): Draw or propagate $Z_{n+1}^{(i)} \sim p(Z_{n+1}|Z_n^{(i)}, y_{n+1})$.
3. (*Estimation*): Draw $\theta^{(i)} \sim p(\theta|Z_{n+1}^{(i)})$.

To do this, we further assume that the posterior predictive $p(y_{n+1}|Z_n)$ and the conditional posterior $p(Z_{n+1}|Z_n, y_{n+1})$ are also available for evaluation. In our examples, we show how to explicitly construct Z_n and the corresponding distributions necessary to implement our approach.

To solve the filtering problem for Z_n we exploit, by construction, the availability of: (i) *Posterior predictive*: $p(y_{n+1}|Z_n, \theta)$ and (ii) *Posterior propagation*: $Z_{n+1} \sim p(Z_{n+1}|Z_n, y_{n+1})$. Reversing the logic of Bayes rule links $p(Z_n|y^n)$ to $p(Z_{n+1}|y^{n+1})$ via

$$p(Z_{n+1}|y^{n+1}) = \int p(Z_{n+1}|Z_n, y^{n+1})\, dP(Z_n|y^n).$$

We can therefore first resample the current particles with the posterior predictive and then propagate the augmented variables. This leads to a simple resample-propagate algorithm for Bayesian learning as detailed in Panel A.

A number of points emerge. First, the predictive distribution $p(y_{n+1}|Z_n)$ will typically depend on only a lower-dimensional subset of components of Z_n, although we will use it to resample the full particle set. Second, information from incoming data is used to build a more efficient set of particles that are then propagated, also informed by y_{n+1}. This has an important practical effect even for pure filtering problems, but the advantage of resampling-first is most pronounced in learning problems where the particles include a set of unknown model parameters. Third, consistency of our estimator is straightforward and we can estimate posterior functionals $E(f(\theta)|y^{n+1}) \approx N^{-1} \sum_{i=1}^N E(f(\theta)|Z_{n+1}^{(i)})$. Finally, an advantage of the resample-propagate framework is that many nonlinearities in the evolution of Z_n are straightforward to incorporate (see Carvalho et al. (2010a); Lopes et al. (2011)).

Constructing Z_n. One novel aspect of our approach is that Z_n can include functions of the parameter θ. To our knowledge, the only similar approach for the construction of Z_n is the nonsequential method of West (1993). The definition of Z_n is not unique. A number of points emerge to find efficient choices. First, as we are just interested in parameter inference, in many cases Z_n^θ will be of fixed

dimension. For example, in hierarchical models a natural set of augmentation variables Z_n corresponds to conditional sufficient statistics in a set of Gibbs complete conditionals.

One property is the existence of a propagation rule for Z_n. This will generally be a combination of a deterministic and/or stochastic propagation rule which we denote

$$Z_{n+1} = p(Z_n, \theta, y_{n+1})$$

given the new data y_{n+1}.

In the state filtering and learning literature, it has been common to use sufficient statistics s_n and parameters θ as augmented variables. In this special case, we can write $Z_n^\theta = (s_n, \theta)$ where s_n is a vector of sufficient statistics that are only dependent on hidden states and data; see Fearnhead (2002) and Storvik (2002) for further discussion. The propagation rule $Z_{n+1} = p(Z_n, \theta, y_{n+1})$ then has two components: first, the deterministic update of sufficient statistics and then a stochastic update of parameters given the sufficient statistic update which can be summarised as

$$s_{n+1} = S(s_n, y_{n+1}) \quad \text{and} \quad \theta^{(i)} \sim p(\theta|s_{n+1}).$$

Another related approach is the missing data algorithm in Kong, Liu and Wong (1994). Here $Z_n = (z_1, \ldots, Z_n)$ tracks the full vector of hidden (missing) variables and sequential importance sampling is used to approximate the joint posterior distribution $p(z_1, \ldots, z_n|y^n)$ and then, as in our approach, Rao-Blackwellisation is used to find $p(\theta|y^n)$. The key difference is that we only track Z_n^θ—a parameter-dependent, fixed-dimensional conditional sufficient statistic. For us to provide the full-joint posterior $p(z_1, \ldots, z_n|y^n)$ we would have to use an extra MCMC step using $p(z_1, \ldots, z_n|\theta^{(i)}, y^n)$ given our parameter draws $\theta^{(i)} \sim p(\theta|y^n)$.

Discussion. It is useful to compare our approach to the current particle filtering/sequential importance sampling literature. Our update for the augmented vector Z_n's can be viewed as a fully adapted auxiliary particle filter (APF, Pitt and Shephard (1999)) with the additional step that the augmented variables can depend on functionals of the parameter (APF is a pure state filtering strategy). The additional parameter draw $\theta^{(i)} \sim p(\theta|Z_n^{(i)})$ is not present in the APF and is used to replenish the diversity of the parameter particles.

In the terminology of the APF, we do not have any second-stage weights as our particles are resampled first and then propagated providing an exact draw from the particle approximation. This has a number of advantages, particularly in the parameter learning context, as we do not introduce any extra variance into the particle weights which can easily lead to particle degeneracies. Moreover, if one carefully defines Z_n to include appropriate auxiliary variables, the flexibility of our approach is not impaired. Examples of this flexibility appears in Carvalho et al. (2010b) in the context of general mixture models.

In the context of state-space models, Storvik (2002) proposes the use of sufficient statistics that are independent of parameters in a propagate-resampling algorithm. Finally, Chen and Liu (2000) work with a similar approach in the mixtures Kalman filters context. Our method differs from the latter in two important ways: (i) they only consider the problem of state filtering and (ii) they work on the propagate-resample framework. This is carefully discussed in Carvalho et al. (2010a) where both Storvik (2002) and Chen and Liu (2000) are extended to the problem of sequential parameter learning in a APF like algorithm for a general class of state-space models. Again, our view of augmented variables Z_n is more general than Storvik's approach. One should view Z_n as more of a computational tool than a probabilistic property of the model such as sufficient statistics.

Another related class of algorithms are Rao-Blackwellised particle filters. The difference is that they typically use propagate-resample algorithm for $Z_t = \{x_{t+1}, z_{t+1}\}$ where z_{t+1} denotes missing data and x_{t+1} a state. As there is no dependence in Z_t on parameters this is a pure filtering problem with parameters estimated offline. Additionally they attempt the approximation of the joint distribution $p(Z^t|y^t)$. This target increases in dimensionality as new data becomes available leading to unbalanced weights. In our framework, $p(Z^t|y^t)$ is not of interest as the filtered, lower-dimensional $p(Z_t|y^t)$ is sufficient for inference at time t. Notice that, based on their work, one has to consider the question of "when to resample?" as an alternative to rebalance the approximation weights. In contrast, our approach requires resampling at every step as the preselection of particles in light of new observations is fundamental in avoiding a decay in the particle approximation for θ.

Another avenue of research uses MCMC steps inside a sequential Monte Carlo algorithm as in the resample-move algorithm of Gilks and Berzuini (2001). This is not required in our strategy as we are using a fully adapted approach. Finally, our general strategies include Liu and West (2001) which gain in their generality but suffer from having to specify tuning parameters.

In order to illustrate the efficiency gains available with our approach consider the most common class of applications: mixture or latent variable models $p(y|\theta) = \int p(y|\lambda)p(\lambda|\theta)\,d\lambda$ where $\lambda^n = (\lambda_1, \ldots, \lambda_n)$ is a data augmentation variable. For this model, with a conditionally conjugate prior, we can find a conditional sufficient statistic, s_n, for parameter learning. Therefore, we define our auxiliary particle variable as $Z_n = (\lambda_n, s_n)$. Under these assumptions, we can write

$$p(\theta|\lambda^{n+1}, y^{n+1}) \sim p(\theta|s_{n+1}) \qquad \text{with } s_{n+1} = \mathcal{S}(s_n, \lambda_{n+1}, y_{n+1}),$$

where $\mathcal{S}(\cdot)$ is deterministic recursion relating the previous s_n to the next, conditionally on λ_{n+1} and y_{n+1}. Now, the propagation step becomes

$$\lambda_{n+1} \sim p(\lambda_{n+1}|\lambda_n, \theta, y_{n+1}),$$
$$s_{n+1} = \mathcal{S}(s_n, \lambda_{n+1}, y_{n+1}).$$

The marginal predictive is given by $p(y_{n+1}|y^n) = \int p(y_{n+1}|\lambda_{n+1}, \theta)p(\lambda_{n+1}, \theta|y^n)\,d\lambda_{n+1}\,d\theta$ which can be easily approximated with our particle draws. See

Kitagawa (1996) for a discussion of marginal likelihood calculation within a state space framework. This avoids the curse of dimensionality encountered if one tries to directly approximate the marginal likelihood by marginalising over (λ^n, θ). The trade-off is that our particle approach will have accumulation of Monte Carlo error and we are approximating $p(y)$ by a product of lower-dimensional integrals.

Marginal likelihoods. Our approach can also provide estimates of the predictive $p(y_{n+1}|y^n)$ and marginal likelihood $p(y^n)$. Marginal likelihoods can then be used to define Bayes factors, the central element to Bayesian model assessment. Following our resampling–sampling approach, an online estimate of the full marginal likelihood can be developed by sequentially approximating $p(y_{n+1}|y^n)$. Specifically, given the current particle draws, we have

$$p(y_{n+1}|y^n) \approx \sum_{i=1}^{N} p\big(y_{n+1}|Z_n^{(i)}\big)$$

which in turn gives us an estimate of the marginal likelihood

$$p(y^n) \approx \prod_{i=1}^{n} p^N(y_i|y^{i-1}).$$

We are therefore able to simplify the problem of calculating $p(y^n)$ by estimating a sequence of small integrals. This also provides access to sequential Bayes factors necessary in many sequential decision problems.

Choice of priors. At its simplest level the algorithm only requires samples $\theta^{(i)}$ from the prior $p(\theta)$. Hence the method is not directly applicable to improper priors. However, the natural class of priors are mixture priors of the form $p(\theta) = \int p(\theta|Z_0)\,p(Z_0)\,dZ_0$. The conditional $p(\theta|Z_0)$ is chosen to be naturally conjugate to the likelihood. If Z_0 is fixed, then we start all particles out with the same Z_0 value. More commonly, we will start with a sample $Z_0^{(i)} \sim p(Z_0)$ and let the algorithm resample these draws with the marginal likelihood $p(y_1|Z_0^{(i)})$. This approach will lead to efficiency gains over blindly sampling from the prior. This method also allows us to implement nonconjugate priors together with vague "uninformative" priors such as Cauchy priors via a scale mixtures of normals.

Monte Carlo error. Due to the sequential Monte Carlo nature of the algorithm, error bounds of the form $C_n/\sqrt{N}$ are available where N is the number of particles used. The constant C_n is model, prior and data dependent and in general its magnitude accumulates over n; see, for example, Brockwell, Del Moral and Doucet (2010). Clearly, these propagated errors will be worse for diffuse priors and for large signal-to-noise ratios as with many Monte Carlo approaches. To assess Monte Carlo standard errors we propose the convergence diagnostic of Carpenter,

Clifford and Fearnhead (1999). By running the algorithm M independent times (based on N particles) one can calculate the Monte Carlo estimates of the mean and variance for the functional of interest. Then by performing an analysis of variance between replicates, the Monte carlo error or effective sample size can be assessed. One might also wish to perform this measure over different data trajectories as some data realizations might be harder to estimate than others. See Lopes et al. (2011) for further discussion on Monte Carlo error.

3 Examples

We now illustrate our approach on two canonical examples. First, the normals-means and ANOVA models presented by Gelfand and Smith (1990) when describing the Gibbs sampling and second, a Bayesian version to regularised lasso regression.

3.1 Normal-means model

Consider a normal-means hierarchical model as a benchmark example. Before the use of simulation-based methods the researcher relied on sophisticated analytical approximations (Tiao and Tan (1965)) or numerical integration. Specifically, assuming that $y_j = (y_{j1}, \ldots, y_{jn_j})'$, we have conditional distributions

$$(y_j|\mu, \theta_j, \sigma^2) \sim N\big((\mu + \theta_j)1_{n_j}, \sigma^2 I_{n_j}\big) \quad \text{and} \quad (\theta_j|\tau^2) \sim N(0, \tau^2),$$

where $j = 1, \ldots, J$, and J is the number of groups, 1_n is an n-dimensional vector of ones and I_n is the identity matrix of order n. Each group j has n_j observations in y_j, treatment level θ_j and overall mean μ. The parameters of interest are μ and τ^2, whose independent prior distributions are $N(\mu_0, \sigma_0^2)$ and $\text{IG}(a_2, b_2)$, where IG denotes an inverse Gamma distribution with mean a_2/b_2. The prior distribution for the nuisance parameter σ^2 is $\text{IG}(a_1, b_1)$ and independent of μ and τ^2.

To construct our augmented variables, we track a vector of parameter dependent set Z_J of the form

$$Z_J = \left(\bar{y}, \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2, \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \mu - \theta_j)^2, \sum_{j=1}^{J}\theta_j^2\right),$$

where y_{ij} is the actual observation. This includes all the necessary sums of squares and satisfies a standard set of recursions. Notice that in this case, the augmented vector is of fixed dimension. Tracking instead $Z_J = (\theta_1, \ldots, \theta_J)$ would lead to more accumulation error as the dimensionality increases with J.

Hence given a new draw θ_{J+1} and data y_{J+1} we can calculate Z_{J+1}. Notice that Z_J depends on parameters such as the overall mean μ as well as individual effects θ_j.

The full conditional posterior distributions of θ_j and μ are

$$(\theta_j|Z_J) \sim N\big(\omega\bar{y} + (1-\omega)\mu 1_J, \omega^2 I_J\big),$$
$$(\mu|Z_J) \sim N\bigg(\sigma_1^2\Big(\sum_{i,j}(y_{ij}-\theta_j)/\sigma^2 + \mu_0/\tau_0^2\Big), \sigma_1^2\bigg),$$

where $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_J)'$, $\omega = J\tau^2/(J\tau^2+\sigma^2)$, $\omega^2 = \sigma^2\tau^2/(J\tau^2+\sigma^2)$, $\sigma_1^{-2} = n\sigma^{-2} + \sigma_0^{-2}$ and $n = \sum_j n_j$. The full conditional posterior distributions of σ^2 and τ^2 are

$$(\sigma^2|Z_J) \sim \mathrm{IG}\bigg(a_1 + n/2, b_1 + \sum_{i,j}(y_{ij}-\mu-\theta_j)^2/2\bigg),$$
$$(\tau^2|Z_J) \sim \mathrm{IG}\bigg(a_2 + J/2, b_2 + \sum_j \theta_j^2/2\bigg).$$

Despite being straightforward to implement, the Gibbs sampler can be trapped in local modes, that is, a small value for τ^2, the random effects variance, will likely lead to small values for θ_j, which in turn will lead to a small value for τ^2 and so forth (see Gelman et al. (2008)). In practice, one would use a collapsed Gibbs sampler (Liu (1994)) and marginalise over the θ_j vector to avoid this problem.

The predictive distribution for the n_{J+1} observations and the posterior distribution for the new random effects θ_{J+1} are

$$(y_{J+1}|Z_J, \gamma) \sim N(\mu 1_{n_{J+1}}, \sigma^2 I_{n_{J+1}} + \tau^2 1_{n_{J+1}} 1'_{n_{J+1}}),$$
$$(\theta_{J+1}|y_{J+1}, \gamma) \sim N\big(C_{J+1} 1'_{n_{J+1}}(y_{J+1} - \mu 1_{n_{J+1}}), C_{J+1}\big),$$

where $\gamma = (\mu, \sigma^2, \tau^2)$ and $C_{J+1}^{-1} = n_{J+1}\sigma^{-2} + \sigma_\theta^{-1}$. Given the particle set $(Z_0, \gamma)^{(i)}$, the resample-propagate algorithm cycles through the following steps:

1. Resample particles with weights $w_{J+1}^{(i)} \propto p(y_{J+1}|Z_J^{(i)}, \gamma^{(i)})$, and form new particle set;
2. Propagate $\theta_{J+1}^{(i)} \sim p(\theta_{J+1}|y_{J+1}, \tilde{\gamma}^{(i)})$;
3. Update conditional sufficient statistics;
4. Draw the components of $\gamma^{(i)}$ from the above full conditional distributions;
5. Derive $Z_{J+1}^{(i)} = \mathcal{S}(Z_J^{(i)}, \gamma^{(i)}, \theta_{J+1}^{(i)})$.

We generated $J = 20$ groups of observations, with $n_j = 5$ replicates per group, and parameter values $\mu = 0$, $\tau^2 = (0.01)^2$ and $\sigma^2 = 1$. Hyper-parameters are set to $\mu_0 = 0$, $\sigma_0^2 = 10^{20}$, $a_1 = b_1 = 0$ and $a_2 = b_2 = 2$. Table 1 and Figure 1 show the comparisons and results based on 300 simulated datasets. The Gibbs sampler gets stuck in local modes very early in the Markov chain simulations, while our resample-propagate algorithm is fairly stable.

Table 1 *Estimation result from* 300 *simulated datasets. Our resample–sample* (*RS*) *scheme uses a Rao-Blackwellized estimator. True values are used for the initial sample of Gibbs sampling. RS is based on* $N = 1000$ *and Gibbs is based on* $N = 10{,}000$

	μ			τ^2		
Method	RMSE	Avg. bias	SE(Avg. bias)	RMSE	Avg. bias	SE(Avg. bias)
RS	0.0986	0.0020	0.0057	0.000134	0.000013	0.000007
Gibbs	0.1086	−0.0064	0.0063	0.021706	0.007654	0.001175

3.2 Sequential Bayesian lasso

We can develop a sequential version of Bayesian lasso (Carlin and Polson (1991) and Hans (2009)) for a simple problem of signal detection. The model takes the form

$$y_i = \theta_i + \varepsilon_i,$$
$$\theta_i = \tau\sqrt{\lambda_i}\varepsilon_i^{\theta},$$

where $\varepsilon_i \sim N(0,1)$, $\varepsilon_i^{\theta} \sim N(0,1)$, $\lambda_i \sim \text{Exp}(2)$ and $\tau^2 \sim \text{IG}(a_0, b_0)$. This leads to independent double exponential marginal prior distributions for each θ_i with $p(\theta_i) = (2\tau)^{-1}\exp(-|\theta_i|/\tau)$. The natural set of latent variables is given by the augmentation variable λ_{n+1} and conditional sufficient statistics leading to $Z_n = (\lambda_{n+1}, a_n, b_n)$. The sequence of variables λ_{n+1} are i.i.d. and so can be propagated directly with $p(\lambda_{n+1})$, whilst the conditional sufficient statistics (a_{n+1}, b_{n+1}) are deterministically determined based on parameters $(\theta_{n+1}, \lambda_{n+1})$ and previous values (a_n, b_n).

Given the particle set $(Z_0, \tau)^{(i)}$, the resample-propagate algorithm cycles through the following steps:

1. Resample particles with weights $w_{n+1}^{(i)} \propto p(y_{n+1}; 0, 1 + \tau^{2(i)}\lambda_{n+1}^{(i)})$, and form new particle set;
2. Propagate $\theta_{n+1}^{(i)} \sim N(m_n^{(i)}, C_n^{(i)})$, $m_n^{(i)} = C_n^{(i)}\tilde{\tau}^{2(i)}\tilde{\lambda}_{n+1}^{(i)}y_{n+1}$ and $C_n^{-1} = 1 + \tilde{\tau}^{-2(i)}\tilde{\lambda}_{n+1}^{-1(i)}$;
3. Update sufficient statistics $a_{n+1}^{(i)} = \tilde{a}_n^{(i)} + 1/2$ and $b_{n+1} = \tilde{b}_n^{(i)} + \theta_{n+1}^{2(i)}/(2\tilde{\lambda}_{n+1}^{(i)})$;
4. Draw $\tau^{2(i)} \sim \text{IG}(a_{n+1}, b_{n+1})$ and $\lambda_{n+2}^{(i)} \sim \text{Exp}(2)$;
5. Let $Z_{n+1}^{(i)} = (\lambda_{n+1}^{(i)}, a_n^{(i)}, b_n^{(i)})$ and update $(Z_{n+1}, \tau)^{(i)}$.

We use our marginal likelihood (or Bayes factor) to compare lasso with a standard normal prior. Under the normal prior we assume that $\tau^2 \sim \text{IG}(a_1, b_1)$ and we match the variances of the parameter θ_i. As the lasso is a model for sparsity

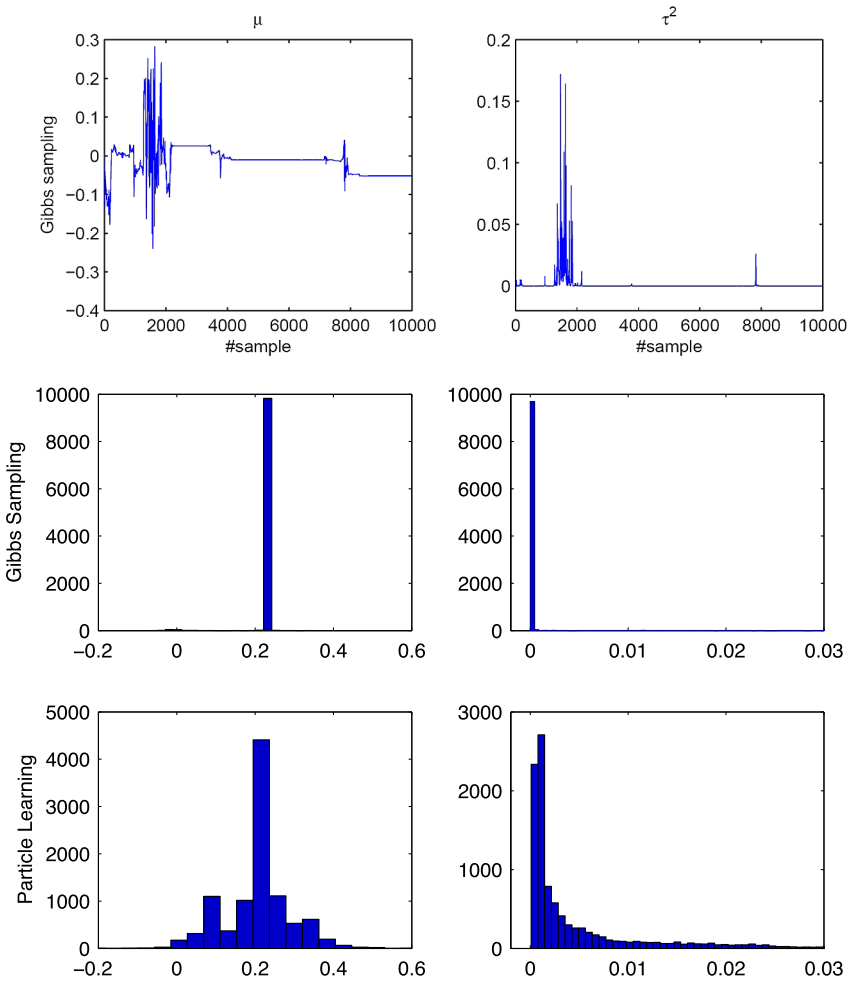

Figure 1 *Top panel: Typical pattern of the trace plots in Gibbs sampling. The samples are stuck for a long time at the value very close to zero. Middle panel: Approximate posterior distributions for the parameters* μ *and* τ^2 (*based on top frame Gibbs draws*). *Bottom panel: Approximate posterior distributions of the parameters* μ *and* τ^2 (*based on our resampling–sampling filter*). *Particle filter and Gibbs sampler sizes are both* 10,000 *draws*.

we would expect the evidence for it to increase when we observe $y_t = 0$. We can sequentially estimate $p(y_{n+1}|y^n, \text{lasso})$ via

$$p(y_{n+1}|y^n, \text{lasso}) = \frac{1}{N}\sum_{i=1}^{N} p\big(y_{n+1}|(\lambda_n, \tau)^{(i)}\big)$$

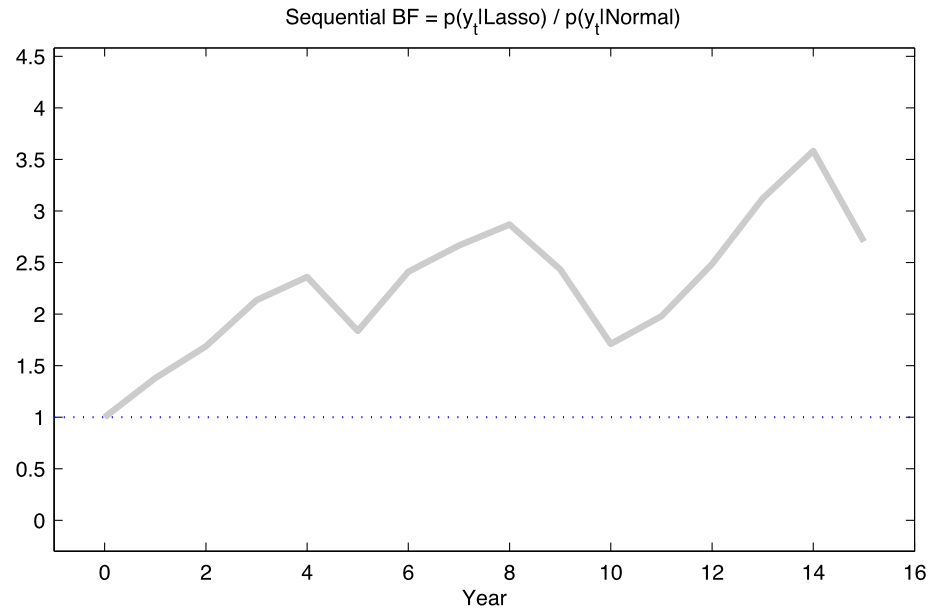

Figure 2 *Sequential Bayes factor*: *Lasso versus normal*. *Particle size is* 10,000.

with predictive $p(y_{n+1}|\lambda_n, \tau) \sim N(0, \tau^2\lambda_n + 1)$. This leads to a sequential Bayes factor

$$BF_{n+1} = \frac{p(y^{n+1}|\text{lasso})}{p(y^{n+1}|\text{normal})}.$$

Data was simulated based on $\theta = (0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1)$ and priors $\tau^2 \sim \text{IG}(2, 1)$ for the double exponential case and $\tau^2 \sim \text{IG}(2, 3)$ for the normal case, reflecting the ratio of variances between those two distributions. Results are summarized by Figure 2 which plots the sequential Bayes factor. As expected the evidence in favor of the lasso is increased when we observe $y = 0$ and for the normal model when we observe a signal $y = 1$.

Our approach can easily be extended to a lasso regression setting. Suppose that we have $y_{t+1} = X_t'\beta + \sigma\sqrt{\lambda_{t+1}}\epsilon_{t+1}$ and $\theta = (\beta, \sigma^2)$ and conditionally conjugate prior is assumed, that is, $p(\beta|\sigma^2) \sim N(b_0, \sigma^2 B_0^{-1})$ and $p(\sigma^2) \sim \text{IG}(\nu_0/2, d_0/2)$. We track $Z_t = (s_t, \lambda_{t+1})$ where $s_t = (b_t, B_t, d_t)$ are conditional sufficient statistics for the parameters. The recursive definitions are $B_{t+1} = B_t + \lambda_{t+1}^{-1} X_t' X_t$, $B_{t+1}b_{t+1} = B_t b_t + \lambda_{t+1}^{-1} X_t' y_{t+1}$ and $d_{t+1} = d_t + b_t' B_t b_t + \lambda_{t+1}^{-1} X_{t+1}' y_{t+1} - b_{t+1}' B_{t+1} b_{t+1}$. The conditional posterior $p(\theta|Z_n)$ is then available for sampling and our approach applies.

We use this example to compare the accuracy in estimating the posterior distribution of the regularization penalty $p(\tau|y)$. We use the generic resample-move

batch importance sampling developed by Gilks and Berzuini (2001) and Chopin (2002). The data is cut into batches parameterized by block-lengths (n, p). In the generic resample move algorithm, we first initialize by drawing from the prior $\pi(\theta, \tau)$ with $\theta = (\theta_1, \ldots, \theta_{15})$. The particles are then resampled with the likelihood from the first batch of observations $(y_1, \ldots, y_p)$. Then the algorithm proceeds sequentially.

There is no need to use the λ_i augmentation variables as this algorithm does not exploit this conditioning information. Then a MCMC kernel is used to move particles. Here, we use a simple random walk MCMC step. This can clearly be tuned to provide better performance although this detracts from the "black-box" nature of this approach. Chopin (2002) provides recommendations for the choice of kernel. Figure 3 provides the comparison with two separate runs of the algorithm both with $N = 10{,}000$ particles for $(n, p) = (3, 5)$ or $(n, p) = (15, 1)$. The performance is similar for the case $p = 1$. Our efficiency gains come from the extra conditioning information available in Z_n.

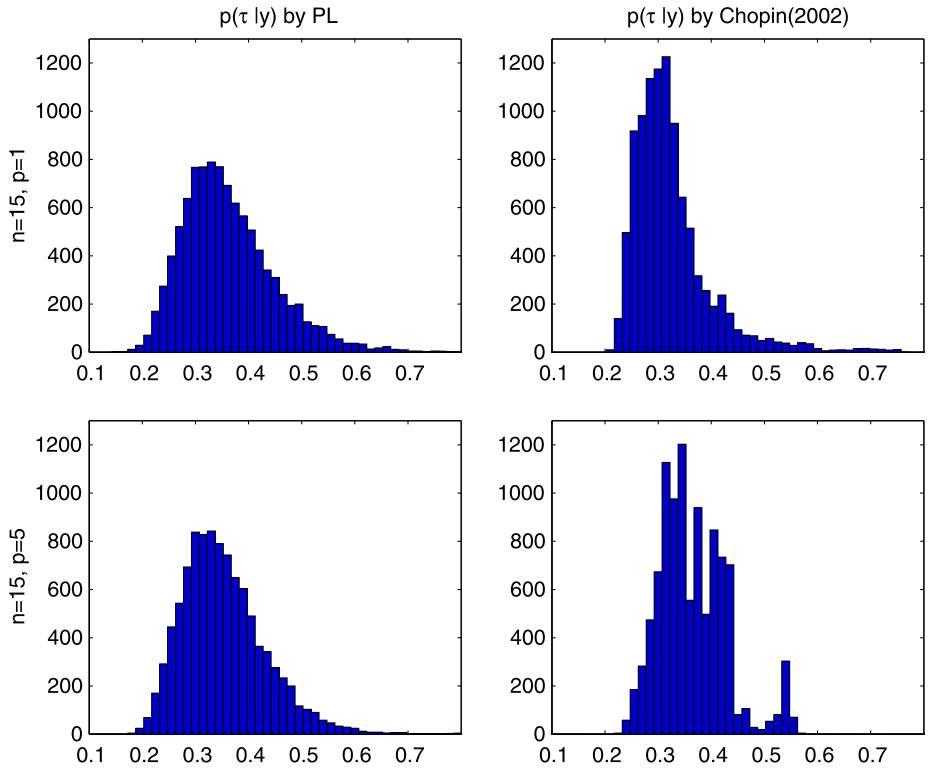

Figure 3 *Comparison to Chopin's (2002) batch sampling scheme.*

4 Final remarks

Our resample–sample perspective is intuitive and easy to implement providing a simple simulation approach to Bayesian statistics. Many of the models that are amenable to MCMC fall into the class where it is easy to construct data augmentation variables to also implement our approach. In hierarchical models we illustrated the simplicity and efficiency of our approach over plain vanilla MCMC algorithms (Gelfand and Smith (1990)) whilst avoiding diagnosing convergence of a Markov chain. As a byproduct, we provide fully sequential inference via the set of posterior distributions $p(\theta|y^n)$. The caveat is that one needs enough probabilistic structure to be able to construct an efficient augmented variable Z_n. In the class of hierarchical models, there are many such choices—motivated by the complete conditional structure for the parameters as in Gibbs sampling.

Possibly the greatest advantage of the methodology is for sequential learning problems where it can be applied to general state space models and where it extends the Kalman filter recursions to learning (Carvalho et al. (2010a); Lopes and Tsay (2011)), to general mixture models including infinite dimensional mixtures such as described by Dirichlet processes (Carvalho et al. (2010b)) and to the sequential Bayesian computation of more general classes of models (Lopes et al. (2011)).

Acknowledgments

We would like to thank the Statistical and Applied Mathematical Sciences Institute (SAMSI) for research support during 2008–2009 Program on Sequential Monte Carlo Methods and the participants of the Particle Learning Working Group. We would also like to thank Seung-Min Yae for research assistance.

References

Brockwell, A., Del Moral, P. and Doucet, A. (2010). Sequentially interacting Markov chain Monte Carlo. *The Annals of Statistics* **38**, 3387–3411. MR2766856

Carlin, B. P. and Polson, N. G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *The Canadian Journal of Statistics* **19**, 399–405. MR1166846

Carpenter, J., Clifford, P. and Fearnhead, P. (1999). An improved particle filter for non-linear problems. *IEE Proceedings—Radar, Sonar and Navigation* **146**, 2–7.

Carvalho, C. M., Johannes, M., Lopes, H. F. and Polson, N. G. (2010a). Particle learning and smoothing. *Statistical Science* **25**, 88–106. MR2741816

Carvalho, C. M., Lopes, H. F., Polson, N. G. and Taddy, M. (2010b). Particle learning for general mixtures. *Bayesian Analysis* **5**, 709–740.

Chen, R. and Liu, J. S. (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society, Series B* **62**, 493–508. MR1772411

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika* **89**, 539–551. MR1929161

Fearnhead, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *Journal of Computational and Graphical Statistics* **11**, 848–862. MR1951601

Gelfand, A. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409. MR1141740

Gelman, A., van Dyk, D. A., Huang, Z. and Biscardin, W. J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics* **17**, 95–122. MR2424797

Gilks, W. and Berzuini, C. (2001). Following a moving target: Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society, Series B* **63**, 127–146. MR1811995

Hans, C. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–845. MR2564494

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian non-linear state space models. *Journal of Computational and Graphical Statistics* **5**, 1–25. MR1380850

Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* **89**, 278–288.

Liu, J. (1994). The collapsed Gibbs sampler with applications to a gene regulation problem. *Journal of the American Statistical Association* **89**, 958–966. MR1294740

Liu, J. and West, M. (2001). Combined parameters and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* (A. Doucet, N. de Freitas and N. Gordon, eds.). New York: Springer. MR1847793

Lopes, H. F., Carvalho, C. M., Johannes, M. S. and Polson, N. G. (2011). Particle learning for sequential Bayesian computation (with discussion). In *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 317–360. Oxford: Oxford Univ. Press.

Lopes, H. F. and Tsay, R. S. (2011). Particle filters and bayesian inference in financial econometrics. *Journal of Forecasting* **30**, 168–209.

Pitt, M. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association* **94**, 590–599. MR1702328

Rubin, D. B. (1987). A noniterative sampling-importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association* **82**, 543–546.

Smith, A. F. M. and Gelfand, A. (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician* **46**, 84–88. MR1165566

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Transactions of Signal Processing* **50**, 281–289.

Tiao, G. C. and Tan, Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. *Biometrika* **52**, 37–54. MR0208765

West, M. (1993). Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society, Series B* **55**, 409–422. MR1224405

H. F. Lopes
N. G. Polson
The University of Chicago Booth
School of Business
5807 S. Woodlawn Ave.
Chicago, IL 60637
USA
E-mail: hlopes@chicagobooth.edu
ngp@chicagobooth.edu

C. M. Carvalho
McCombs School of Business
University of Texas at Austin
B6500, CBA 5.202, University Station
Austin, TX 78712
USA
E-mail: carlos.carvalho@mccombs.utexas.edu