An efficient computational approach for prior sensitivity analysis and cross-validation

Luke BORNN*, Arnaud DOUCET and Raphael GOTTARDO

Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC, Canada V6T 1Z2

Key words and phrases: Cross-validation; g-prior; Markov chain Monte Carlo; penalized regression; prior sensitivity; regularization path plot; sequential Monte Carlo; variable selection.

MSC 2000: Primary 62F15; secondary 65C05.

Abstract: Prior sensitivity analysis and cross-validation are important tools in Bayesian statistics. However, due to the computational expense of implementing existing methods, these techniques are rarely used. In this paper, the authors show how it is possible to use sequential Monte Carlo methods to create an efficient and automated algorithm to perform these tasks. They apply the algorithm to the computation of regularization path plots and to assess the sensitivity of the tuning parameter in g-prior model selection. They then demonstrate the algorithm in a cross-validation context and use it to select the shrinkage parameter in Bayesian regression. *The Canadian Journal of Statistics* © 2010 Statistical Society of Canada

Résumé: La sensibilité à la loi a priori et la validation croisée sont des outils importants des statistiques bayésiennes. Toutefois, ces techniques sont rarement utilisées en pratique car les méthodes disponibles pour les implémenter sont numériquement très coûteuses. Dans ce papier, les auteurs montrent comment il est possible d'utiliser les méthodes de Monte Carlo séquentielles pour obtenir un algorithme efficace et automatique pour implémenter ces techniques. Ils appliquent cet algorithme au calcul des chemins de régularisation pour un problème de régression et à la sensibilité du paramètre de la loi a priori de Zellner pour un problème de sélection de variables. Ils appliquent ensuite cet algorithme pour la validation croisée et l'utilisent afin de sélectionner le paramètre de régularisation dans un problème de régression bayésienne. *La revue canadienne de statistique* © 2010 Société statistique du Canada

1. INTRODUCTION AND MOTIVATION

An important step in any Bayesian analysis is to assess the prior distribution's influence on the final inference. In order to check prior sensitivity, the posterior distribution must be studied using a variety of prior distributions. If these posteriors are not available analytically, they are usually approximated using Markov chain Monte Carlo (MCMC) methods. Since obtaining the posterior distribution for one given prior can be very expensive computationally, repeating the process for a large range of prior distributions is often prohibitive. Importance sampling has been implemented as an attempted solution (Besag et al., 1995), but the potential of infinite variance importance weights makes this technique useless if the posterior distribution changes more than a trivial amount as the prior is altered. Additionally, this importance weight degeneracy typically increases with the dimension of the parameter space.

One such prior sensitivity problem is the computation of regularization path plots—a commonly used tool when performing penalized regression. In these situations there is typically a tuning parameter which controls the amount of shrinkage on the regression coefficients; regularization path plots graphically display this shrinkage as a function of the tuning parameter.

[*] *Author to whom correspondence may be addressed.*
E-mail: l.bornn@stat.ubc.ca

More precisely, these plots display the shrunken coefficients as a function of their L_1-norm in order to facilitate comparison between competing shrinkage methods. This choice of norm originates from the LASSO shrinkage and variable selection method of Tibshirani (1996), for which the LARS algorithm (Efron et al., 2004) may be employed to quickly produce these plots. In the Bayesian version (Vidakovic, 1998; Park & Casella, 2008), however, we may want to plot the posterior means (or other posterior summary statistics) of the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ for a range of the tuning (or penalty) parameter λ. Using a double exponential prior on $\boldsymbol{\beta}$, the corresponding posterior distributions are proportional to

$$\exp\left(-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\lambda\sum_{j=1}^{p}|\beta_j|\right]\right) \tag{1}$$

where the response $\mathbf{y}$ is assumed to come from a normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and variance σ^2 for a model matrix $\mathbf{X}$. Since approximating (1) using MCMC at one level of λ can take upwards of an hour depending on the precision required, producing this plot by repeating MCMC hundreds of times for different λ is impractical.

Another tool requiring repeated posterior approximations is cross-validation, which has two primary statistical purposes. The first is finding the value of a given parameter (for instance, the penalty parameter in penalized regression) which minimizes prediction error. The second is comparing different models' or methodologies' prediction performance. In both situations the data are split into a training set, which is used to fit the model, and a testing set, which is used to gauge the prediction performance of the trained model. A typical example would involve fitting a model on the training set for a range of values of some model parameter, then setting this parameter to the value that results in the lowest prediction error rate on the testing set. For example, we might wish to select the value of λ in (1) to minimize prediction error. From a computational standpoint, cross-validation is similar to prior sensitivity in both structure and complexity. Further, the entire process is usually repeated for a variety of different training and testing sets and the results are then combined. Although importance sampling has been applied to cross-validation (e.g., Alqallaf & Gustafson, 2001), the problem of infinite variance importance weights remains (Peruggia, 1997).

In this paper, we begin by motivating and developing sequential Monte Carlo (SMC) methods, then subsequently apply them to prior sensitivity analysis and cross-validation. While the SMC methods themselves are well-established in the literature, to the authors' knowledge their application to prior sensitivity and cross-validation has yet to be explored. In Section 2 we present an efficient algorithm for sampling from a sequence of potentially quite similar probability distributions defined on a common space. Section 3 demonstrates the algorithm in a prior sensitivity setting and applies it to the creation of regularization path plots and the sensitivity of the tuning parameters when performing variable selection using g-priors. Cross-validation with application to Bayesian regression is developed in Section 4. We close with extensions and concluding remarks in Section 5.

2. SEQUENTIAL MONTE CARLO ALGORITHMS

SMC methods are often used in the analysis of dynamic systems where we are interested in approximating a sequence of probability distributions $\pi_t(\theta_t)$ where $t = 1, 2, 3, \ldots, T$. The variable θ_t can be of evolving or static dimension as t changes; note that t is simply an index variable and need not be real time. Most work in the SMC literature is interested in the evolving dimension case, with applications to state-space models (Doucet, Godsill & Andrieu, 2000) and target tracking (Liu & Chen, 1998) among others. The static case, where each π_t lies in a common space, has received less attention (Chopin, 2002; Del Moral, Doucet & Jasra, 2006). The goal of SMC methods is to sample from the distributions $\{\pi_t\}$ sequentially, that is, first from π_1, then π_2, up to π_T. In some situations we are concerned with each intermediate distribution, whereas in others only the final distribution π_T is of interest (e.g., Neal, 2001). For further reading, the edited volume of Doucet, de Freitas & Gordon (2001) covers a range of developments in SMC theory and applications.

The situation where the sequence of distributions lies in a common space arises in several applications. For instance, the number of observations in some experiments can make MCMC prohibitive. In this case π_t might be the posterior distribution of a parameter given the observations 1 through t. Moving through the data with a sequential strategy in this way may decrease computational complexity. Another application is transitioning from a simple distribution π_1 to a more complex distribution of interest π_T. Alternatively we could consider situations analogous to simulated annealing (Kirkpatrick, Gelatt & Vecchi, 1983), where $\pi_t(\theta) \propto [\pi(\theta)]^{\phi_t}$ for an increasing sequence $\{\phi_t\}, t = 1, 2, 3, \ldots, T$. In all of these examples the bridging distributions $\pi_1, \ldots, \pi_{T-1}$ are only used to reach the final distribution of interest π_T. When we are interested in a certain feature of each π_t, SMC will typically be computationally cheaper than MCMC even if we can successfully sample from each π_t using MCMC. This is because SMC borrows information from adjacent distributions, using the samples from earlier distributions to help in approximating later distributions. Often the difficulty in using SMC is constructing this sequence of distributions; both prior sensitivity and cross-validation are situations where there exists a natural sequence upon which SMC may be applied. From here forward we assume the distributions to have a common support.

For all times t, we seek to obtain a collection of N weighted samples (called particles) $\{W_t^{(i)}, \theta_t^{(i)}\}, i = 1, \ldots, N$ approximating π_t where the weights are positive and normalized to sum to 1. We may estimate expected values with these particles using $\hat{E}_{\pi_t}(g(\theta)) = \sum_{i=1}^N W_t^{(i)} g(\theta_t^{(i)})$. One technique used in SMC is importance sampling, where particles $\{W_{t-1}^{(i)}, \theta_{t-1}^{(i)}\}$ distributed as π_{t-1} may be reused, reweighting them (before normalization) according to

$$W_t^{(i)} \propto W_{t-1}^{(i)} \frac{\pi_t(\theta_{t-1}^{(i)})}{\pi_{t-1}(\theta_{t-1}^{(i)})} \tag{2}$$

in order to obtain an approximation of π_t. Thus we obtain the current weights by multiplying the previous weights by an incremental weight $\pi_t(\theta_{t-1}^{(i)})/\pi_{t-1}(\theta_{t-1}^{(i)})$.

In an attempt to prevent these weights from becoming overly non-uniform, we may move each particle $\theta_{t-1}^{(i)}$ (currently distributed according to π_{t-1}) with a Markov kernel $K_t(\theta, \theta')$ to a new position $\theta_t^{(i)}$, then subsequently reweight the moved particles to be distributed according to π_t. Although the kernel $K_t(\theta, \theta') = \pi_t(\theta')$ minimizes the variance of the importance weights, it is typically impossible to sample from; thus it has been proposed to use Markov kernels with invariant distribution π_t (Gilks & Berzuini, 2001). A direct application of this strategy suffers from a major flaw, however, as the importance distribution given by

$$\eta_t(\theta_t) = \int \pi_1(\theta_1) \prod_{t=2}^{T} K_t(\theta_{t-1}, \theta_t)\, \mathrm{d}\theta_{1:T-1}$$

is usually impossible to compute and therefore we cannot calculate the necessary importance weights. Additionally, this assumes we are able to sample from $\pi_1(\theta_1)$, which is not always the case. Alternatives attempt to approximate η_t pointwise when possible, but the computation of these algorithms is in $O(N^2)$ (Del Moral et al., 2006).

The central idea of SMC samplers (Del Moral et al., 2006) is to employ an auxiliary backward kernel with density $L_{t-1}(\theta_t, \theta_{t-1})$ to get around this untractable integral.

It can be interpreted as a generalization of the method presented in Crooks (1998) and Neal (2001). This backward kernel relates to a time-reversed SMC sampler giving the same marginal distribution as the forward SMC sampler induced by $K_t(\theta_{t-1}, \theta_t)$. The backward kernel is essentially arbitrary, but should be optimized to minimize the variance of the importance weights. Del Moral et al. (2006) prove that the sequence of backward kernels minimizing the variance of the importance weights is, for any t, $L_{t-1}^{\mathrm{opt}}(\theta_t, \theta_{t-1}) = \eta_{t-1}(\theta_{t-1})K_t(\theta_{t-1}, \theta_t)/\eta_t(\theta_t)$. However, it is typically impossible to use this optimal kernel since it relies on intractable marginals. Thus, we should select a backward kernel that approximates this optimal kernel. Two suboptimal backward kernels given in Del Moral et al. (2006) to approximate L_{t-1}^{opt} are $\pi_{t-1}(\theta_t)K_t(\theta_{t-1}, \theta_t)/(\pi_{t-1}K_t(\theta_t))$ and $\pi_t(\theta_{t-1})K_t(\theta_{t-1}, \theta_t)/\pi_t(\theta_t)$. The latter is the same as that used explicitly in Crooks (1998) and Neal (2001) and implicitly in Chopin (2002) and Gilks & Berzuini (2001). These two backward kernels result in respective incremental weights

$$w_t^a(\theta_{t-1}, \theta_t) = \frac{\pi_t(\theta_t)}{\int \pi_{t-1}(\theta_{t-1})K_t(\theta_{t-1}, \theta_t)\,\mathrm{d}\theta_{t-1}} \tag{3a}$$

$$w_t^b(\theta_{t-1}, \theta_t) = \frac{\pi_t(\theta_{t-1})}{\pi_{t-1}(\theta_{t-1})}. \tag{3b}$$

These incremental weights are then multiplied by the weights at the previous time and normalized to sum to 1. We note that the suboptimal kernel resulting in (3b) can be thought of as an approximation of that resulting in (3a), and has the same form as (2), the reweighting mechanism for importance sampling. Despite this, both kernels lead to a correct algorithm. In this manner the first kernel should perform better, particularly when successive distributions are considerably different (Del Moral et al., 2006). Although the weights (3a) are a better approximation of the optimal backward kernel weights, the second kernel is convenient since the resulting incremental weights (3b) do not depend on the position of the moved particles θ_t and hence we are able to reweight the particles prior to moving them. We include the incremental weight (3a) because, when K_t is a Gibbs kernel (i.e., $K_t(\theta_{t-1}, \mathrm{d}\theta_t) = \delta_{\theta_{t-1,-k}}(\mathrm{d}\theta_{t,-k})\pi_t(\mathrm{d}\theta_{t,k}|\theta_{t,-k})$) moving one subset k at a time, it simplifies to $\pi_t(\theta_{t-1,-k})/\pi_{t-1}(\theta_{t-1,-k})$ where $\theta_{t-1,-k}$ is the particle excluding the kth component. By a simple Rao–Blackwell argument it can be seen that this choice, by conditioning on the variable being moved, results in reduced variance of the importance weights compared to (3b).

2.1. An Efficient SMC Algorithm

Now that we have described some components of SMC methodology, we proceed to develop an efficient algorithm for performing prior sensitivity and cross-validation. The basic idea of our algorithm is to first reweight the particles $\{W_{t-1}^{(i)}, \theta_{t-1}^{(i)}\}$, $i = 1, \ldots, N$, such that they are approximately distributed as π_t. If the variance of the weights is large, we then resample the particles with probabilities proportional to their weights, giving us a set of N equally weighted particles (including some duplicates). After resampling we move the particles with a kernel of invariant distribution π_t, which creates diversity in the particles. Our algorithm relates closely to resample-move algorithms (Gilks & Berzuini, 2001; Chopin, 2002), although our formulation is more general and allows for the use of a variety of suboptimal backward kernels and corresponding weights.

Moving the particles at each time step is not particularly efficient. For example, if two successive distributions in the sequence are identical, we are wasting our time by moving the particles. If successive distributions are similar but not necessarily identical, to save computational time we can simply copy forward the particles at time $t-1$ and reweight them with the importance sampling weights (2). Deciding when to move particles may be done dynamically or deterministically. A dynamic scheme would move the particles whenever the variance of the weights becomes too large (usually measured by the effective sample size (ESS) $(\sum_{i=1}^{N}(W_t^{(i)})^2)^{-1}$), whereas a deterministic scheme would move the particles every kth time step for some integer k. Since the sequence of distributions will likely not change at a constant rate, it is better to use a dynamic scheme as this allows for little particle movement during parts of the sequence with little change and more movement in parts of the sequence where successive distributions vary more. The frequency of particle movement should be partially determined by knowledge of the mixing properties of the Markov kernel used. For instance, if the kernel mixes very well, we can move particles less frequently, whereas slow-mixing kernels require more frequent movement to ensure particles adequately follow the sequence of distributions.

When the ESS drops below a specified threshold, we reweight the particles at time $t-1$ to be approximately distributed as π_t prior to moving them. The weights (3b) only depend on the particles at time $t-1$, so we can easily do this. In the case of a one at a time Gibbs sampler, we can also use the weights (3a). Because the unweighted particles at time t are not distributed according to π_t, we cannot simply move the particles without first taking their weights into consideration. Thus prior to moving the particles we resample them such that $W_t^{(i)} = 1/N$ for $i = 1, \ldots, N$ and the particles' unweighted distribution is π_t. Resampling methods duplicate particles with large weights and remove particles with low weights. Specifically, we copy the ith particle $N_t^{(i)}$ times such that $\sum_{i=1}^{N} N_t^{(i)} = N$ and $E(N_t^{(i)}) = NW_t^{(i)}$, where $W_t^{(i)}$ are the normalized importance weights. Lastly, all of the resampled particles are assigned equal weights. The simplest unbiased resampling method consists of sampling $N_t^{(i)}$ from a multinomial distribution with parameter $(N, \{W_t^{(i)}\})$. It should be noted that more sophisticated resampling schemes, such as residual resampling (Liu, 2001) and stratified resampling (Kitagawa, 1996) exist, resulting in reduced variance of $N_t^{(i)}$ relative to multinomial resampling. After the particles are resampled, we can move them with the kernel K_t.

An efficient SMC algorithm which may be used to perform prior sensitivity and cross-validation is therefore:

```
for $t = 1$ do
    Obtain $N$ weighted samples $\theta_1^{(i)}$ from $\pi_1$ (directly, MCMC, etc.)
end
for $t = 2, \ldots, T$ do
    Copy $\theta_{t-1}^{(i)}$ to $\theta_*^{(i)}$ and calculate weights $W_*^{(i)} \propto W_{t-1}^{(i)} \times \frac{\pi_t(\theta_{t-1}^{(i)})}{\pi_{t-1}(\theta_{t-1}^{(i)})}$
    if $ESS(W_*) > c$ then
        Copy $\left(\theta_*^{(i)}, W_*^{(i)}\right)$ to $\left(\theta_t^{(i)}, W_t^{(i)}\right)$
    else
        Move: Move particles with Markov kernel of invariant distribution $\pi_t$,
        $\theta_t^{(i)} \sim K_t(\theta_{t-1}^{(i)}, \cdot)$
        Reweight: Calculate weights according to $W_t^{(i)} \propto W_{t-1}^{(i)} \times w_t(\theta_{t-1}^{(i)}, \theta_t^{(i)})$ where
        $w_t(\theta_{t-1}^{(i)}, \theta_t^{(i)})$ is either given by (3a) or (3b)
        Resample: Resample particles according to above weights. Set all weights to $1/N$
    end
end
```

note 1: If a backward kernel is chosen such that the incremental weights are independent of the moved particle $\theta_t^{(i)}$, e.g. (3b), the reweight and resample step come before the move step.
note 2: c is a user-specified threshold on the effective sample size.

3. PRIOR SENSITIVITY

In the case of prior sensitivity we are interested in approximating the posterior distribution of some variable(s) θ given the data D, symbolically notated as $\pi(\theta|D) \propto f(D|\theta)\nu(\theta)$, where $f(D|\theta)$ and $\nu(\theta)$ are the likelihood and the prior distribution of θ, respectively. Here the notation $\nu(\theta)$ is used to differentiate the prior from the posterior distribution $\pi_t(\theta)$, allowing for the omitting of dependencies. This prior sensitivity framework has been studied in a closed-form setting (Gustafson & Wasserman, 1995; Gustafson, 1996), but situations requiring Monte Carlo methods have received less attention. It is worth noting that only the prior distribution changes between successive distributions (the likelihood remains the same). Thus when we reweight particles to approximate the sequence of posterior distributions for θ, the weights (2) depend solely on the prior distribution,

$$\begin{aligned} W_t^{(i)} &\propto W_{t-1}^{(i)} \frac{f(D|\theta_{t-1}^{(i)})\nu_t(\theta_{t-1}^{(i)})}{f(D|\theta_{t-1}^{(i)})\nu_{t-1}(\theta_{t-1}^{(i)})} \\ &\propto W_{t-1}^{(i)} \frac{\nu_t(\theta_{t-1}^{(i)})}{\nu_{t-1}(\theta_{t-1}^{(i)})} \end{aligned} \quad (4)$$

where $\theta_t^{(i)}$ is the ith particle sampled at time t and $\nu_t(\theta_t^{(i)})$ is the tth prior distribution evaluated at the point $\theta_t^{(i)}$. If the ESS falls below a given threshold at time t (notated as c in algorithm pseudocode), we resample and move, otherwise we simply reweight. Conveniently, resampling and moving using (3b) and reweighting using (2) both result in the same weight mechanism (4). In a later example we will also employ the weights (3a), which have reduced variance relative to (3b).

3.1. Regularization Path Plots

Consider the regression model with response vector $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and model matrix $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p})$ where $\mathbf{x_j} = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$, $j = 1, \ldots, p$ are the column vectors of predictors (including the unit intercept vector). For clarity of presentation we present the model with a continuous response; however, it is simple to extend to binary responses (Albert & Chib, 1993). We use the prostate data of Stamey et al. (1989) which has eight predictors and a response (logarithm of prostate-specific antigen) with likelihood

$$\mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n). \quad (5)$$

Using a double exponential prior distribution with parameter λ on the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, the corresponding posterior distribution is proportional to (1). We see from the form of this posterior distribution that if $\lambda = 0$ the MAP estimate of $\boldsymbol{\beta}$ will correspond to the least-squares solution. However, as λ increases there will be shrinkage on $\boldsymbol{\beta}$ which may be displayed using a regularization path plot. Because the shrinkage as λ varies is nonlinear, we set a schedule $\lambda_t = \mathrm{e}^{t/20}, t = 1, \ldots, 100$. We create a "gold standard" Bayesian Lasso regularization path plot for this data by running MCMC with a Markov chain of length 100,000 at each level of λ and plotting the posterior mean of the resulting regression coefficients (Figure 1). Because accurately estimating extreme quantiles using MCMC is difficult, we also show a 99% credible interval for two of the coefficients. It should be noted that the creation of this plot took over 8 h.

Since the idea is to create these plots quickly for exploratory analysis, we will compare our SMC-based method to MCMC with both constrained to work in 5 min (±5 s), and both using the same Markov kernel. In order to perform MCMC in our time frame of 5 min, the Markov chain had a length of 1,200 for each of the 100 levels of λ. To maximize efficiency, the final state of each Markov chain was used as the initial state of the subsequent chain. The mean of each resulting posterior distribution was used to plot the regularization path plot in Figure 2a. In comparison, to

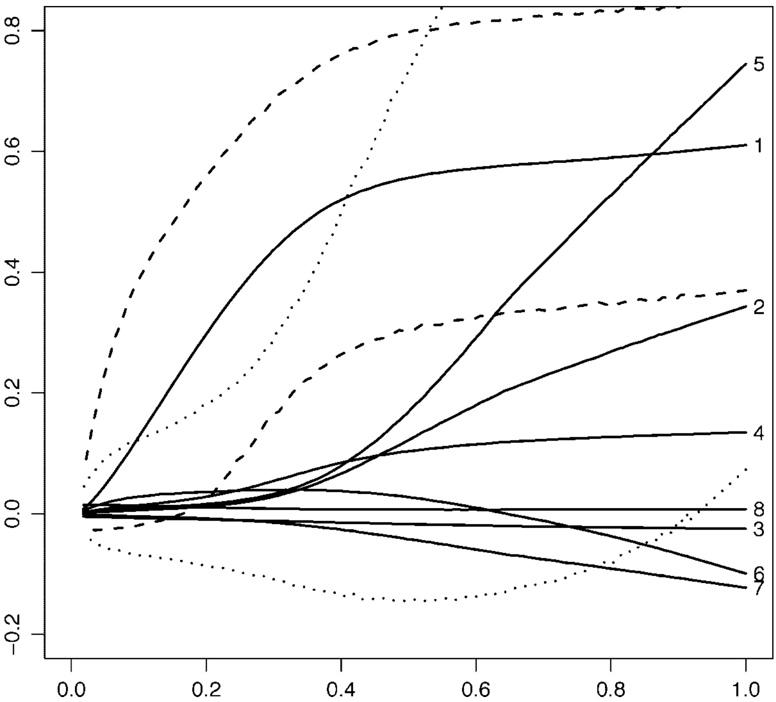

FIGURE 1: Regularization Path Plots: The gold standard. The plot is of standardized coefficients β_j versus $|\boldsymbol{\beta}|_1/\max(|\boldsymbol{\beta}|_1)$. Ninety-nine percent credible intervals for coefficients 1 and 5 shown with dashed and dotted lines, respectively.

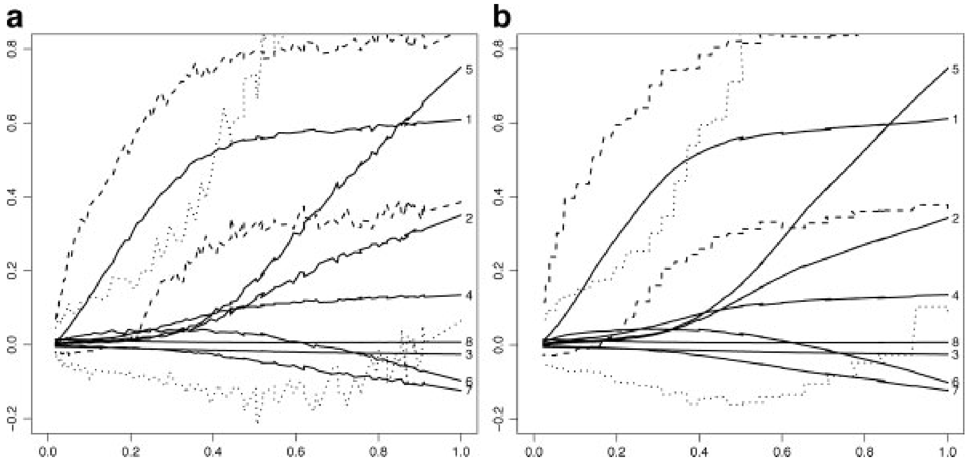

FIGURE 2: Regularization path plots: Plots using MCMC and SMC for fixed computational time of 5 min. The plots are of standardized coefficients β_j versus $|\boldsymbol{\beta}|_1/\max(|\boldsymbol{\beta}|_1)$. Ninety-nine percent credible intervals for coefficients 1 and 5 shown with dashed and dotted lines, respectively. (a) MCMC with 1,200 samples (5 min); (b) SMC with 4,200 samples (5 min).

run in 5 min our SMC algorithm used $N = 4{,}200$ particles and resampled and moved the particles when the ESS dropped below $c = 2N/3 = 2{,}800$ (Figure 2b). For the sake of time comparisons, all computations here and later were performed on a Power Mac G5 with dual 2.7 GHz PowerPC processors. We see from these plots that both methods capture the key features of the regularization path plot as shown by the gold standard: every one of the variables has the correct path. The methods vary, however, in the amount of noise. In order to estimate the extreme quantiles, both algorithms used the corresponding sample quantiles, and hence because of the larger sample size, the SMC algorithm provides better estimates. In estimating the mean, we see much more variability using MCMC compared to SMC. This is due to SMC being able to use many more particles since it is able to save time by borrowing information from previous distributions. To be specific, the SMC algorithm in this context had to resample and move the particles only 25 times in the entire sequence of 100 distributions. The remainder of the time our algorithm simply reweighted the particles, which is computationally inexpensive. This feature is largely responsible for the smooth results of the SMC algorithm. Hence it is possible that if our initial sample is biased, subsequent reweightings might also be biased. However, assuming the Markov kernels K_t are mixing fast enough, any initial error will be forgotten in later distributions (Del Moral, 2004, Chapter 7). In subsequent examples we explore this bias-variance trade-off for each algorithm. It is worth noting that, because of this reweighting mechanism, adding more incremental distributions in the sequence will have little effect on the computational time of the SMC algorithm, unlike MCMC-based strategies, which would approximate each new distribution with a new Markov chain. In addition, we attempted to make these plots using importance sampling, reweighting (and not moving) particles from π_1 to approximate later distributions. However, the weights became degenerate, with all of the weights eventually focussing on one particle with standardized L1 norm of 0.8. Specifically, all but one of the weights had values near zero, and the one particle with positive weight had standardized L1 norm of 0.8. Thus importance sampling was only able to create roughly 1/5 of the entire plot, and hence is clearly not a candidate methodology for creating these plots. We will see later that in many such situations importance sampling fails, even with large amounts of particles.

3.2. Variable Selection Using g-Priors

Consider the normal likelihood set-up (5). Now, however, with an eye towards model selection, we introduce the binary indicator variable $\gamma \in \{0, 1\}^p$, where $\gamma_j = 1$ means the variable $\boldsymbol{x}_j$ is included in the model. Thus γ can describe all of the 2^p possible models. Following the notation of Marin & Robert (2007), we use $q_\gamma = \mathbf{1}_n^{\mathrm{T}}\gamma$ as a counter for the number of variables in the model. If $\mathbf{X}_\gamma$ is the model matrix which excludes all $\boldsymbol{x}_j$'s if $\gamma_j = 0$, we can employ the following prior distributions for $\boldsymbol{\beta}$ and σ^2 (Zellner, 1986; Marin & Robert, 2007):

$$\pi(\boldsymbol{\beta}_\gamma, \sigma^2|\gamma) \propto (\sigma^2)^{-(q_\gamma+1)/2-1} \exp\left[-\frac{1}{2g\sigma^2}\boldsymbol{\beta}_\gamma^{\mathrm{T}}(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)\boldsymbol{\beta}_\gamma\right].$$

From this it is straightforward to show that the posterior density for γ is thus

$$\pi(\gamma\,|\,\mathbf{y}, \mathbf{X}) \propto (g+1)^{-(q_\gamma+1)/2}\left[\mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{g}{g+1}\mathbf{y}^{\mathrm{T}}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma\mathbf{y}\right]^{-n/2}. \qquad (6)$$

We perform model selection on the pollution data set of McDonald & Schwing (1973), in which mortality rate is compared against 15 pollution-related variables in 60 metropolitan areas. The 15 independent variables include, for instance, mean annual precipitation, population per household, and average annual relative humidity. The response variable $\mathbf{y}$ is the age-adjusted mortality rate in the given metropolitan area. We seek to perform variable selection to narrow down the number of independent variables which best predict the response. With 15 variables, calculating the posterior probabilities of the over 30,000 models exactly is possible but time-consuming. We have chosen this size of data set to allow for a benchmark from which we can compare MCMC to SMC.

Our goal is to see how the explanatory variables change as we vary the prior distribution parameter g. In other words, we are interested in seeing how robust the variable selection method is to changes in the setting of g, specifically $g = \mathrm{e}^{t/10}$, $t = 1, \ldots, 100$. We use a Gibbs sampler strategy to compare the SMC-based algorithm to brute-force MCMC, benchmarked against the exact solution obtained from (6), in which β_γ and σ^2 are integrated out. Specifically, we update γ one component at a time. The incremental weight ratio (3b) will be the ratio of the posterior distribution (6) evaluated on the complete data at successive levels of g. In addition, we are able to use the weights (3a), which corresponds to the ratio of the posterior distribution (6) evaluated on all of the data, excluding the variable that is being moved by the Gibbs sampler.

In order to see our desired result, we use (6) to plot the exact marginal probabilities as well as some sample model probabilities for various levels of g (Figure 3a and b). The models chosen are those with highest posterior probability in this range of g. This process took slightly over 8 h, and hence we would like to find a faster method. We constrain both stochastic algorithms to run in 30 min (±1 min). As a result the MCMC algorithm uses a Markov chain of length 10,000 and the SMC algorithm uses 18,000 particles. At each time step, a randomly chosen variable from each particle is added/removed from the model. We plot the resulting posterior marginal and model probabilities for each algorithm in Figure 3c–f. First impression shows that the plot created using MCMC has much more variability. However, the smoothness in the SMC algorithm is not a result of perfect accuracy of the method, but rather only smoothness of the reweighting mechanism (2). Because of this, if SMC does poorly during times of particle movement, the subsequent reweighted approximations will also be inaccurate. To ensure this is not the case and verify that SMC is indeed outperforming MCMC, we look at the average absolute error of the marginal probabilities (at 100 levels of λ and for 15 variables). We find the average absolute error in the marginal probabilities using MCMC is 0.0292 whereas with SMC it is only 0.0187. In addition, their respective maximum absolute errors were 0.24 and 0.08, respectively. In fact 50 runs of the algorithms resulted in similar results, with SMC consistently outperforming MCMC. Specifically, the respective mean (and standard deviation) for SMC and MCMC for the average absolute error were 0.0182 (0.0016) and 0.0289 (0.0026), and for the maximum absolute error were 0.11 (0.04) and 0.28 (0.03). From this we see that SMC is indeed providing a better approximation of the true marginal probabilities.

What then may be taken from these marginal probability plots? When performing simple forward selection regression, the variables 1, 2, 6, 9, and 14 are chosen. Slightly different results come from doing backward selection; in particular variables 1 and 14 are replaced by variables 12 and 13. The LASSO solution (using fivefold cross-validation) is the same as the forward solution with the additional variables 7 and 8. In addition, the LASSO solution contains some shrinkage on the regression coefficients (see Example 3.1). Using the marginal probabilities resulting from g-priors, the variables that clearly stand out (see Figure 3a) are 1, 2, 6, 9, and 14. Thus the g-prior solution taken from the plot corresponds to the forward selection model. As g increases, the most probable model increases from $\{9\}$ to $\{6, 9\}$ to $\{2, 6, 9\}$. Also, for a given g, say $g = \mathrm{e}^9$, the marginal probability plot obtained with SMC shows the correct top 4 variables for inclusion, whereas the variability from the MCMC-based plot makes it impossible to do so.

4. CROSS-VALIDATION

We focus on leave-s-out cross-validation, which is the case when the testing set consists of s observations. Continuing in the linear regression framework, let $\mathbf{X}_{\backslash S}$ and $\mathbf{y}_{\backslash S}$ be the model matrix and response vector excluding the subset S of observations (of size s). We are interested

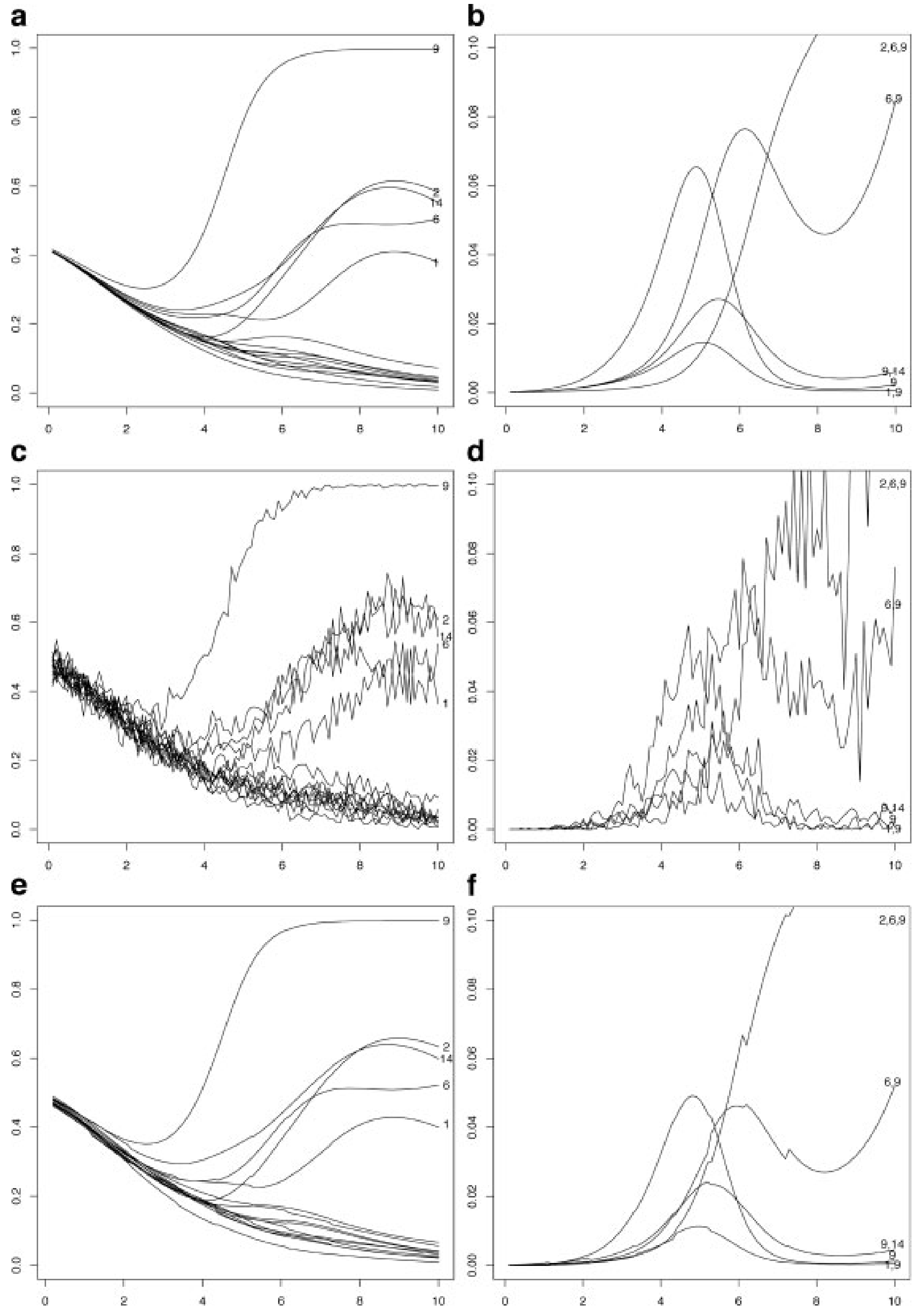

FIGURE 3: Marginal and model probabilities for variable selection using g-priors as a function of $\log(g)$. Plot (a) highlights several variables ($\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_6$, $\mathbf{X}_9$, $\mathbf{X}_{14}$) which show high marginal probabilities of inclusion. Plot (b) shows the posterior probabilities of five models chosen to highlight the effect of g on model size. Plots (c–f) compare MCMC to SMC's performance. (a) Posterior marginal probabilities: Exact solution. (b) Posterior model probabilities: Exact solution. (c) Posterior marginal probabilities: MCMC with 10,000 samples (30 min). (d) Posterior model probabilities: MCMC with 10,000 samples (30 min). (e) Posterior marginal probabilities: SMC with 18,000 samples (30 min). (f) Posterior model probabilities: SMC with 18,000 samples (30 min).

in a collection of T model parameter (typically prior distribution parameter) settings resulting in posterior densities $\pi_t(\theta|\mathbf{X}_{\backslash S}, \mathbf{y}_{\backslash S})$ for $t = 1, \ldots, T$. Once we have approximations of all T posterior densities, we select the model parameter settings which result in the best prediction of $\mathbf{y}_S$ using $\mathbf{X}_S$. To find the sequence of distributions $\pi_t(\theta|\mathbf{X}_{\backslash S}, \mathbf{y}_{\backslash S})$, $t = 1, \ldots, T$, the same SMC-based algorithm proposed for prior sensitivity is applicable. Specifically, once we have obtained a Monte Carlo approximation of $\pi_1(\theta|\mathbf{X}_{\backslash S}, \mathbf{y}_{\backslash S})$, we can transition to the remainder of the distributions $\pi_t(\theta|\mathbf{X}_{\backslash S}, \mathbf{y}_{\backslash S})$, $t = 2, \ldots, T$ using SMC.

In addition to quickly evaluating the model for a variety of settings on the training set, SMC methods also provide a tool for switching the training/testing set without fully re-approximating the posterior densities. Specifically, suppose we have a testing set S_1, and using SMC we find approximations of $\pi_t(\theta|\mathbf{X}_{\backslash S_1}, \mathbf{y}_{\backslash S_1})$, $t = 1, \ldots, T$, each of which are tested for prediction performance on the subset S_1. However, typically we are interested in performing cross-validation for a variety of different splits of the data into training and testing sets. Thus, we will now want a new testing set S_2 and find approximations of $\pi_t(\theta|\mathbf{X}_{\backslash S_2}, \mathbf{y}_{\backslash S_2})$, $t = 1, \ldots, T$. The obvious way to accomplish this is to start fresh by approximating $\pi_1(\theta|\mathbf{X}_{\backslash S_2}, \mathbf{y}_{\backslash S_2})$ with MCMC and proceeding to approximate the remainder of the distributions using SMC. However, we can be a bit more clever than this, recognizing that $\pi_1(\theta|\mathbf{X}_{\backslash S_1}, \mathbf{y}_{\backslash S_1})$ and $\pi_1(\theta|\mathbf{X}_{\backslash S_2}, \mathbf{y}_{\backslash S_2})$ are related (Alqallaf & Gustafson, 2001; Bhattacharya & Haslett, 2007).

Successive splits of the data into training and testing sets should give similar model settings. Therefore, we first build the model for a given parameter setting on the full data set using SMC, resulting in an approximation of $\pi_1(\theta|\mathbf{X}, \mathbf{y})$. Then instead of using MCMC to get approximations of $\pi_1(\theta|\mathbf{X}_{\backslash S}, \mathbf{y}_{\backslash S})$ for different $S \in \{S_1, \ldots, S_{\max}\}$, we can build a sequence of distributions $(\pi_1(\theta|\mathbf{X}, \mathbf{y}))^{1-\gamma}(\pi_1(\theta|\mathbf{X}_{\backslash S}, \mathbf{y}_{\backslash S}))^{\gamma}$ for an increasing temperature $\gamma = 0, \epsilon, 2\epsilon, \ldots, 1 - \epsilon, 1$ which will allow us to transition to the case-deletion posteriors. The process is illustrated in Figure 4. The case of $\gamma = 0, 1$ with no movement step corresponds to basic case-deletion importance sampling as discussed in Peruggia (1997). Although case-deletion importance sampling has been demonstrated to achieve up to 90% cost savings in some circumstances (Alqallaf & Gustafson, 2001), the problem of degeneracy still makes importance sampling fail in many situations (Peruggia, 1997; Epifani, MacEachern & Peruggia, 2005).

$$\begin{tikzcd}
 & \pi_1(\theta|\mathbf{X},\mathbf{y}) \arrow[dl] \arrow[dr] & \\
\pi_1(\theta|\mathbf{X}_{\backslash S_1},\mathbf{y}_{\backslash S_1}) \arrow[d] & \cdots\cdots & \pi_1(\theta|\mathbf{X}_{\backslash S_{max}},\mathbf{y}_{\backslash S_{max}}) \arrow[d] \\
\pi_2(\theta|\mathbf{X}_{\backslash S_1},\mathbf{y}_{\backslash S_1}) \arrow[d, dotted] & \cdots\cdots & \pi_2(\theta|\mathbf{X}_{\backslash S_{max}},\mathbf{y}_{\backslash S_{max}}) \arrow[d, dotted] \\
\pi_T(\theta|\mathbf{X}_{\backslash S_1},\mathbf{y}_{\backslash S_1}) & \cdots\cdots & \pi_T(\theta|\mathbf{X}_{\backslash S_{max}},\mathbf{y}_{\backslash S_{max}})
\end{tikzcd}$$

FIGURE 4: Diagram of cross-validation process. Each arrow represents transitioning using SMC.

Let $\Theta = (\boldsymbol{\beta}, \sigma^2)$. The posterior distribution $\pi(\Theta)$ of Θ is proportional to $q(\Theta) = f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta}) \times \pi(\sigma^2)$. Assume we collect samples from the distribution $\pi(\Theta)$. We are interested in reweighting these samples such that they come from the distribution attained by removing the set S. The modified likelihood and posterior for this case-deletion scenario are, respectively

$$f_{\setminus S}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = (\sigma^2)^{-(n-s)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{y}_S - \mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}((\mathbf{y}_S - \mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}))\right]\right\}$$

$$q_{\setminus S}(\Theta) = f_{\setminus S}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta}) \times \pi(\sigma^2)$$

We assume that the prior distributions for $\boldsymbol{\beta}$ and σ^2 are proper and independent. Epifani et al. (2005) show that if the weights $w_{\setminus S}(\Theta) = q_{\setminus S}(\Theta)/q(\Theta)$ are used to move to the case-deletion posterior directly, then the rth moment of these weights is finite if and only if all of the following conditions hold:

$$\begin{aligned} &(a) \quad \lambda_H < 1/r \\ &(b) \quad n - rs > 1 \\ &(c) \quad \mathrm{RSS}^*{}_{\setminus S}(r) > 0 \end{aligned}$$

where λ_H is the largest eigenvalue of the matrix $H_S = \mathbf{X}_S^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}_S$ and $\mathrm{RSS}^*{}_{\setminus S}(r) = \mathrm{RSS} - re_S^{\mathrm{T}}(\mathbf{I} - rH_S)^{-1}e_S$ where $e_S = y_S - \mathbf{X}_S^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ and RSS denotes the residual sum of squares of the least-squares fit of the full data set. This result should not be taken lightly: as Geweke (1989) points out, if the second moment does not exist, the importance sampling estimator will follow neither a $N^{1/2}$ asymptotic (where N is the number of importance sampling draws) nor a central limit theorem. (a) states that if the leverage of the deleted observations is too large, then the importance weights will have infinite variance. (b) gives a condition relating sample size to the allowable test set size s. (c) says that if the influence of the deleted observation is large relative to RSS, then the importance weights will have infinite variance. We show here how using a sequence of artificial intermediate distributions with SMC can help to mitigate this problem.

We introduce a sequence of distributions

$$q_\gamma(\Theta) \propto (q(\Theta))^{1-\gamma}(q_{\setminus S}(\Theta))^\gamma$$

where $\gamma = 0, \epsilon, 2\epsilon, \ldots, 1 - \epsilon, 1$ to move from $q(\Theta)$ to $q_{\setminus S}(\Theta)$. At a given step $\gamma = \gamma^*$ in the sequence, the successive importance weights appearing in the SMC algorithm to move to the next step $\gamma^* + \epsilon$ are

$$\begin{aligned} w_{\setminus S,\gamma^*}(\Theta) &= \frac{(q(\Theta))^{1-\gamma^*-\epsilon}(q_{\setminus S}(\Theta))^{\gamma^*+\epsilon}}{(q(\Theta))^{1-\gamma^*}(q_{\setminus S}(\Theta))^{\gamma^*}} \\ &= \left(\frac{q_{\setminus S}(\Theta)}{q(\Theta)}\right)^\epsilon \end{aligned}$$

Theorem 1. *Provided that* $\mathrm{RSS}{*}_{\setminus S}(1) > 0$ *and the prior distributions for* $\boldsymbol{\beta}$ *and* σ^2 *are proper and independent, a sequence of distributions proportional to* $\{(q(\Theta))^{1-\gamma}(q_{\setminus S}(\Theta))^\gamma; \gamma = 0, \epsilon, 2\epsilon, \ldots, 1 - \epsilon, 1\}$ *may be constructed to move from* $q(\Theta)$ *to* $q_{\setminus S}(\Theta)$ *such that the importance weights* $w_{\setminus S,\gamma}(\Theta)$ *for each successive step have a finite* rth *moment under* $q_\gamma(\Theta)$ *provided*

$$\epsilon < \frac{\alpha - 1}{r - 1} \tag{7}$$

where $\alpha > 1$ is chosen to satisfy

$$\lambda_H < 1/\alpha \tag{8a}$$

$$n - \alpha s > 2 \tag{8b}$$

$$\mathrm{RSS}{*}_{\setminus S}(\alpha) > 0. \tag{8c}$$

The proof may be found in the Appendix. The provision that $\mathrm{RSS}{*}_{\setminus S}(1) > 0$ is very reasonable, and states that the least-squares fit of the full data must not fit the training set perfectly. Note also that we find α for each subset S. Thus we may use the largest allowable step size ϵ in (7) for each subset S, maximizing the algorithm's efficiency by varying the length of the sequence of tempered distributions for each subset. While this result is not sufficient to establish that the variance of SMC estimates are finite for a finite number N of particles, it can be used to upper bound the asymptotic variance of SMC estimates under additional mild regularity mixing conditions on the MCMC kernels; see Chopin (2004), Del Moral (2004, Chapter 7), and Jasra & Doucet (2008) for similar ideas.

4.1. Application to Bayesian Regression

To demonstrate the strength of SMC applied to cross-validation, we use it to select the parameter λ of the Bayesian Lasso (1). For brevity, we reuse the pollution data set (McDonald & Schwing, 1973) of Section 3.2, selecting the parameter λ using leave-one-out cross-validation. Firstly, it is worth pointing out that importance sampling will fail in this situation, as $\lambda_H > 1/2$ on 6 of the 60 observations in this data set, and hence the sufficient conditions to ensure finite variance are not satisfied. Using a sequence of intermediate distributions, we find that the largest α satisfying (8) equals 1.103, or, to ensure a finite second moment, $\epsilon < (\alpha - 1)/(r - 1) = 0.103$. Thus, it suggests using a sequence of distributions of length at most 10. For most variables $\alpha > 2$, which for $r = 2$ is equivalent to importance sampling. Thus SMC does not waste time transitioning to case-deleted posteriors if importance sampling will suffice.

We use a Gibbs sampler to approximate the posterior distribution of (β, σ^2) for $\lambda = \mathrm{e}^{-5}$ on the full data set and then use SMC to move to the case-deletion posterior distributions by creating a sequence of auxiliary distributions as described above. For each different case-deletion we then use SMC to find approximations of the posterior for schedule $\lambda = \mathrm{e}^{t/15}, t = -75, \ldots, 75$. Plotting the cross-validation errors as a function of λ using MCMC with a Markov chain of length 20,000 (Figure 5, solid line) we observe that the average squared loss $\sum_{k=1}^{21}(\mathrm{y}_k - \boldsymbol{x}_k\boldsymbol{\beta})^2/21$ is a smooth function in λ with minimum near $\mathrm{e}^{3/2}$. This "gold standard" plot required 48 h to complete. Thus to minimize prediction error (at least in terms of the squared loss) we should set $\lambda = \mathrm{e}^{3/2}$. To perform this task in a time-restricted manner we constrained both MCMC and SMC algorithms to work in 30 min (±1 min). Figures 5a and b are the resulting plots. The reduced variability of the SMC-based plot allows us to make more accurate conclusions. For instance, it is clear in the plot obtained with SMC (Figure 5b) that the minimum error lies somewhere around $\lambda = \mathrm{e}^{3/2}$, whereas from the MCMC plot (Figure 5a) it could be anywhere between $\mathrm{e}^{1/2}$ and $\mathrm{e}^{5/2}$.

5. EXTENSIONS AND CONCLUSIONS

In our presentation of the algorithm, a fixed sequence of distributions $\pi_t(\theta), t = 1, 2, 3, \ldots, T$ is used. However, it is also possible to determine the sequence of distributions automatically such that successive distributions are a fixed distance apart, as measured by ESS. For instance, assume we are interested in $\pi_t(\theta) = \pi(\theta|\lambda_t)$ where λ_t is a scalar parameter and we have a Monte Carlo approximation of $\pi(\theta|\lambda_{t-1})$ for an arbitrary t, namely $\{W_{t-1}^{(i)}, \theta_{t-1}^{(i)}\}, i = 1, \ldots, N$. We may set λ_t

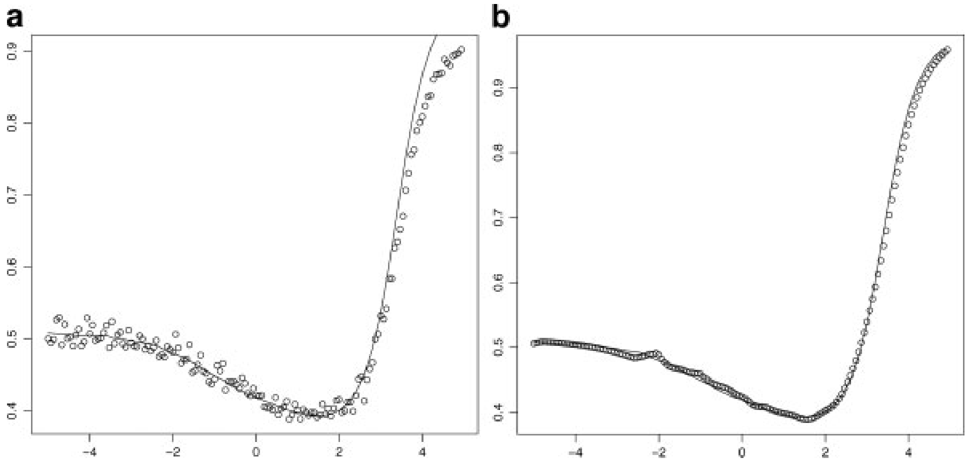

FIGURE 5: Plots of cross-validation error as a function of log(λ). (a) Cross-validation error as a function of log(λ) using MCMC with 60 samples (30 min). Gold standard (MCMC with 20,000 samples) shown with solid line. (b) Cross-validation error as a function of log(λ) using SMC with 350 samples (30 min). Gold standard (MCMC with 20,000 samples) shown with solid line.

to ensure that ESS $= c$ for a constant c by solving

$$c = \sum_{i=1}^{N} \left((W_t^{(i)})^2 \right)^{-1}$$

where $W_t^{(i)}$ is given by (2). This may be solved numerically or in closed-form, if possible. This technique would be beneficial in situations where little or nothing is known about the sequence of distributions, and hence it would be nice to automatically create the sequence.

All our examples have considered a sequence of distributions parameterized by a scalar parameter for which the definition of the sequence of target distributions is very intuitive. If we are interested in dealing with multivariate parameters then the algorithm may be adapted by, for instance, creating a grid (or hyper-grid) of distributions. SMC may be used to work across each dimension in succession. It is worth noting that the complexity of the algorithm scales exponentially with dimension, although MCMC does as well.

Also of interest is the parallelization of SMC algorithms to further decrease the time required to perform prior sensitivity and cross-validation. While recent work has primarily focussed on cluster computing environments, promising progress has been made using graphics processing units or GPUs. See, for instance, Lee et al. (2009), who reduce computational time of SMC methods by upwards of two orders of magnitude by conducting massively parallel inference with GPUs.

While we have given two choices of incremental weights, (3a) and (3b), many other choices are available (Del Moral et al., 2006). In situations where the weights are dependent on the position of the moved particle, such as with (3a), auxiliary particle techniques may be used (Pitt & Shephard, 1999; Johansen & Whiteley, 2009). Specifically, we reweight the particles with an approximation of the weight of interest (for instance, (3a)) which is only dependent on the particles at time $t-1$, using $W_{\text{temp}}^{(i)} \propto W_{t-1}^{(i)} \times W_{\text{approx}}^{(i)}$ where $W_{\text{approx}}^{(i)}$ is the approximation of the incremental weight. After we have resampled and moved the particles we then compensate for this approximation using $W_t^{(i)} \propto W_{\text{true}}^{(i)} / W_{\text{approx}}^{(i)} \times W_{\text{temp}}^{(i)}$ where $W_{\text{true}}^{(i)}$ is the true weights given by (3a) or (3b).

We have seen that by adapting importance sampling to move particles between successive distributions, SMC drastically limits the problem of importance sampling degeneracy. By using a resample-move type algorithm, we are able to perform prior sensitivity and cross-validation in a computationally feasible manner while avoiding the fore-mentioned pitfalls of importance sampling. We have shown the SMC algorithm to be considerably more efficient than existing methods based on iterative MCMC approximations. In this way regularization path plots and other sensitivity analysis problems can be studied in the context of the full posterior distribution instead of a few summary statistics. In addition, SMC provides a tool for naturally performing cross-validation in an efficient manner. Lastly, through the importance weights, SMC provides a measure of the distance between distributions, and hence gives a way to select a subset of distributions of interest for exploratory or other purposes.

APPENDIX

Proof of Theorem 1. (following along the lines of Peruggia (1997) and Epifani et al. (2005)) to show that the rth moment of successive importance weights is finite, we need to find the conditions under which $\int \phi(\Theta)\,\mathrm{d}\Theta$ is finite, where $\phi(\Theta) = (q(\Theta))^{1-\gamma}(q_{\backslash S}(\Theta))^{\gamma} \times (w_{\backslash S,\gamma}(\Theta))^{r}$. We expand and simplify $\phi(\Theta)$ to obtain

$$\begin{aligned}
\phi(\Theta) &= f^{1-\gamma}(\mathbf{y}|\boldsymbol{\beta},\sigma^2) \times f_{\backslash S}^{\gamma}(\mathbf{y}|\boldsymbol{\beta},\sigma^2) \times \pi(\boldsymbol{\beta}) \times \pi(\sigma^2) \times (w_{\backslash S,\gamma}(\Theta))^{r} \\
&= f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) \times [w_{\backslash S,\gamma}(\Theta)]^{\gamma+r\epsilon} \times \pi(\boldsymbol{\beta}) \times \pi(\sigma^2) \\
&= (\sigma^2)^{-((n-s(\gamma+r\epsilon))/2-1)-1} \times \pi(\boldsymbol{\beta}) \times \pi(\sigma^2) \\
&\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - (\gamma+r\epsilon)(\mathbf{y}_S-\mathbf{X}_S\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}_S-\mathbf{X}_S\boldsymbol{\beta})\right]\right\} \\
&= \phi_1(\Theta) \times \phi_2(\Theta)
\end{aligned}$$

where

$$\begin{aligned}
\phi_1(\Theta) &= \pi(\boldsymbol{\beta}) \times \pi(\sigma^2) \times \exp\left\{-\frac{1}{2\sigma^2}[(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S](\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})]\right\} \\
\phi_2(\Theta) &= (\sigma^2)^{-((n-s(\gamma+r\epsilon))/2-1)-1} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma^2}[\mathbf{y}^{\mathrm{T}}\mathbf{y} - (\gamma+r\epsilon)\mathbf{y}_S^{\mathrm{T}}\mathbf{y}_S - \tilde{\boldsymbol{\beta}}^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}}]\tilde{\boldsymbol{\beta}}]\right\}
\end{aligned}$$

and $\tilde{\boldsymbol{\beta}} = [\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}}]^{-1}[\mathbf{y}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{y}_S\mathbf{X}_S^{\mathrm{T}}]$. We will show momentarily that $[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}}]$ is positive definite, and hence invertible. Note that $\phi_1(\Theta)$ is proportional to a proper density for Θ when $[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}}]$ is positive definite. In this case $\phi_1(\Theta)$ is upper bounded. Now $\phi_2(\Theta)$ is proportional to an inverse gamma distribution provided that both

$$\frac{n-s(\gamma+r\epsilon)}{2} > 1$$

$$\mathbf{y}^{\mathrm{T}}\mathbf{y} - (\gamma+r\epsilon)\mathbf{y}_S^{\mathrm{T}}\mathbf{y}_S - \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\left[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\right]\tilde{\boldsymbol{\beta}} > 0$$

Thus, aside from showing conditions under which $[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}}]$ is positive definite, we also need to find conditions guaranteeing the above two inequalities. We first show that $[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma+r\epsilon)\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S]$ is positive definite. Using the Woodbury matrix identity, we see that

$[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma + r\epsilon)\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S]^{-1}$ may be written as

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\gamma + r\epsilon)\mathbf{X}_S(\mathbf{I} - (\gamma + r\epsilon)\mathbf{X}_s^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}_S)^{-1}\mathbf{X}_S^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

Now if $(\mathbf{I} - (\gamma + r\epsilon)\mathbf{X}_s^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}_S)^{-1}$ is positive definite, the second term in the above sum is positive semi-definite. This is the case when all the eigenvalues of $\mathbf{X}_s^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}_S$ are less than $1/(\gamma + r\epsilon)$. $[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma + r\epsilon)\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S]^{-1}$ may then be written as the sum of a positive definite and a positive semi-definite matrix, and hence $[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma + r\epsilon)\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S]$ is positive definite.

Now we proceed to find conditions ensuring

$$\mathbf{y}^{\mathrm{T}}\mathbf{y} - (\gamma + r\epsilon)\mathbf{y}_S^{\mathrm{T}}\mathbf{y}_S - \tilde{\boldsymbol{\beta}}^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma + r\epsilon)\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}}]\tilde{\boldsymbol{\beta}} > 0.$$

Simple but tedious algebra gives the following expression:

$$\begin{aligned}
&\mathbf{y}^{\mathrm{T}}\mathbf{y} - (\gamma + r\epsilon)\mathbf{y}_S^{\mathrm{T}}\mathbf{y}_S - \tilde{\boldsymbol{\beta}}^{\mathrm{T}}[\mathbf{X}^{\mathrm{T}}\mathbf{X} - (\gamma + r\epsilon)\mathbf{X}_S\mathbf{X}_S^{\mathrm{T}}]\tilde{\boldsymbol{\beta}} \\
&\quad = \mathrm{RSS} - (\gamma + r\epsilon)e_S^{\mathrm{T}}(\mathbf{I} - (\gamma + r\epsilon)H_S)e_S \\
&\quad = \mathrm{RSS}^*{}_{\backslash S}(\gamma + r\epsilon)
\end{aligned}$$

which, by the theorem's conditions, is greater than 0 for argument value 1, and since $\mathrm{RSS}^*{}_{\backslash S}$ is a smoothly decreasing function in its argument, it is also positive for some positive argument value less than 1. Now, we choose $\epsilon < (\alpha - 1)/(r - 1)$, which implies $\alpha > \gamma + r\epsilon$. By α satisfying (8), the conditions outlined in the proof hold. Namely, (a) $\lambda_H < 1/(\gamma + r\epsilon)$, since these eigenvalues are upper bounded by 1, (b) $n - s(\gamma + r\epsilon) > 2$, and (c) $\mathrm{RSS}^*{}_{\backslash S}(\gamma + r\epsilon) > 0$. ■

ACKNOWLEDGEMENTS

This work was funded by the Natural Sciences and Engineering Research Council of Canada. The authors are grateful to the editor, associate editor, and anonymous referee whose comments greatly improved the article.

BIBLIOGRAPHY

J. H. Albert & S. Chib (1993). Bayesian analysis of binary and polytomous response data. *Journal of the American Statistical Association*, 88, 669–679.

F. Alqallaf & P. Gustafson (2001). On cross-validation of Bayesian models *Canadian Journal of Statistics*, 29, 333–340.

J. Besag, P. Green, D. Higdon & K. Mengersen (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, 10, 58–66.

S. Bhattacharya & J. Haslett (2007). Importance re-sampling MCMC for cross-validation in inverse problems. *Bayesian Analysis*, 2, 385–408.

N. Chopin (2002). A sequential particle filter method for static models. *Biometrika*, 89, 539–552.

N. Chopin (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of Statistics*, 32, 2385–2411.

G. E. Crooks (1998). Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics*, 90, 1481–1487.

P. Del Moral (2004). "Feynman-Kac formulae: Genealogical and interacting particle systems with applications," Springer, New York.

P. Del Moral, A. Doucet & A. Jasra (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68, 411–436.

A. Doucet, N. de Freitas & N. J. Gordon, Editors, (2001). “Sequential Monte Carlo Methods in Practice,” Springer, New York.

A. Doucet, S. Godsill & C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10, 197–208.

B. Efron, T. Hastie, I. Johnstone & R. Tibshirani (2004). Least angle regression. *Annals of Statistics*, 32, 407–499.

I. Epifani, S. MacEachern & M. Peruggia (2005). Case-deletion importance sampling estimators: Central limit theorems and related results. *Technical Report No. 720, Department of Statistics, Ohio State University*.

J. Geweke (1989). Bayesian inference in econometric models using Monte Carlo integration. *Journal of the American Statistical Association*, 88, 881–889.

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

P. Gustafson & L. Wasserman (1995). Local sensitivity diagnostics for Bayesian inference. *Annals of Statistics*, 23, 2153–2167.

P. Gustafson (1996). Local sensitivity of inferences to prior marginals. *Journal of the American Statistical Association*, 91, 774–781.

A. Jasra & A. Doucet (2008). Stability of sequential Monte Carlo samplers via the Foster-Lyapunov condition. *Statistics and Probability Letters*, 78, 3062–3069.

A. M. Johansen & N. Whiteley (2009). A Modern perspective on auxiliary particle filters, In *Proceedings of Workshop on Inference and Estimation in Probabilistic Time Series Models*. Issac Newton Institute, June 2008.

S. Kirkpatrick, C. D. Gelatt & M. P. Vecchi (1983). Optimization by simulated annealing. *Science*, 220, 671–680.

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

A. Lee, C. Yau, M. Giles, A. Doucet & C. Holmes (2009). On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods. *arXiv:0905.2441v3*.

J. S. Liu & R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93, 1032–1044.

J. S. Liu (2001). “Monte Carlo Strategies in Scientific Computing,” 2nd ed., Springer, New York.

J. M. Marin & C. P. Robert (2007). Bayesian Care: a practical approach to computational Bayesian statistics. Springer-Verlag, New York.

G. McDonald & R. Schwing (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15, 463–481.

R. Neal (2001). Annealed importance sampling. *Statistics and Computing*, 11, 125–139.

T. Park & G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103, 681–686.

M. Peruggia (1997). On the variability of case-deletion importance sampling weights in the Bayesian Linear Model. *Journal of the American Statistical Association*, 92, 199–207.

M. K. Pitt & N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94, 590–591.

T. A. Stamey, J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E.A. Redwine & N. Yang (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. *Journal of Urology*, 16, 1076–1083.

R. Tibshirani (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267–288.

B. Vidakovic (1998). “Wavelet-Based Nonparametric Bayes Methods.” In *“Practical Nonparametric and Semiparametric Bayesian Statistics.”* D. Dey, P. Muller & D. Sinha (eds). Springer, New York, pp. 133–256.

A. Zellner (1986). On assessing prior distributions and Bayesian regression analysis with G-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 6, 233–243.

Received 7 August 2008
Accepted 23 July 2009