

# Bayesian Computation and the Linear Model

MATTHEW J. HEATON

*Department of Statistical Science*  
*Duke University*  
matt@stat.duke.edu

JAMES G. SCOTT

*Department of Information, Risk, and Operations Management*  
*University of Texas at Austin*  
james.scott@mcombs.utexas.edu

September 2009

## Abstract

This paper is a review of computational strategies for Bayesian shrinkage and variable selection in the linear model. Our focus is less on traditional MCMC methods, which are covered in depth by earlier review papers. Instead, we focus more on recent innovations in stochastic search and adaptive MCMC, along with some comparatively new research on shrinkage priors. One of our conclusions is that true MCMC seems inferior to stochastic search if one's goal is to discover good models, but that stochastic search can result in biased estimates of variable inclusion probabilities. We also find reasons to question the accuracy of inclusion probabilities generated by traditional MCMC on high-dimensional, nonorthogonal problems, though the matter is far from settled.

*Some key words:* adaptive MCMC; linear models; shrinkage priors; stochastic search; variable selection

## 1 INTRODUCTION

The linear model is a venerable topic, and one that may even seem passé in light of the past decade's revolution in applied Bayesian nonparametric modeling. Yet despite its apparent simplicity, the linear model remains as important as ever to the practice of modern Bayesian statistics, for at least three reasons.

First, many data sets are simply too high-dimensional to be modeled using the slickest, newest methods. Computers run out of memory; Markov chains fail to converge; priors become prohibitively difficult to elicit or choose in a default way. Already this is a problem with data that arises in genetics and SNP association studies. Yet these data sets are small compared to those concerning Internet

traffic that, for example, Google or Microsoft encounter every day. When a linear model is all that can be fit, it should be fit using the best available statistical and computational tools.

Second, many practitioners will fit a linear model to their data as a first pass, and will never make, or never be able to publish, a second pass. Indeed, many decisions in public health and policy are made using the results of a linear regression, with choices of great consequence coming down to the question of whether a particular term is “significant.” Echoing the above: when a linear model is all that *will* be fit, it should be fit using the best available statistical and computational tools.

Finally, some nonparametric, nonlinear models can be recast as parametric, linear ones. For example, many kernel-regression problems correspond to expanding a function as a linear combination of basis elements given by the orthonormal eigenfunctions of an integral operator. Similarly, methods based on wavelets, splines, Fourier polynomials, and many other “dictionaries” of basis functions can be treated as little more than linear regression, and yet are capable of fitting highly nonlinear functions. A hypothetical Bayesian who knew only how to fit “ $Y = X\beta + \text{error}$ ” could still handle a vast array of problems, simply by being clever about the choice of  $X$ .

Complicating matters is the fact that Bayesian linear modeling can generate many potential summaries for a high-dimensional data set, and that each summary corresponds, in some sense, to a different inferential goal. These summaries can include posterior means or medians of regression coefficients, variable-inclusion probabilities, and the posterior probabilities of models themselves (where a model is a specific combination of coefficients being identically zero). Section 4, for example, considers a data set where ozone-concentration levels around Los Angeles are regressed upon 65 possible atmospheric predictors. One could ask at least three different, scientifically relevant, questions concerning this data:

1. **Which subset of atmospheric variables best accounts for observed variation in past ozone levels?** This question can, in principle, be answered by finding the model with the highest posterior probability, given the data and prior assumptions.
2. **Which subset of atmospheric variables should be used to predict future ozone levels?** It is known that model-averaged predictions are generally best, but this is unsatisfactory if one must choose a single model to use for prediction. In orthogonal and nested-model settings, the best model to use for prediction is the median probability model (Barbieri and Berger, 2004). But in general, it is unknown whether there exists a single best model to use for prediction.
3. **What numbers should be used to yield the best estimate of the**

**marginal effect of each variable on ozone levels?** Here, as above, the model-averaged estimates of the coefficients are generally best.

As this list suggests, different methodological approaches may work better for different questions. This paper seeks to reach a complementary understanding of different computational strategies. Indeed, we find that no single computational strategy works best for all of these problems—a fact that is both interesting and surprising, given that all strategies are, fundamentally, trying to reconstruct the same joint distribution over data, models, and parameters. In light of this fact, it is important to understand each strategy’s strengths and weaknesses.

Our approach differs from existing review papers on Bayesian linear models in two main ways:

1. We focus less on well-established material regarding traditional MCMC, and more on recent innovations involving stochastic search and adaptive MCMC.
2. We provide a computational and methodological overview of “pure shrinkage” solutions, which have been the subject of a recent surge in research activity. An example of a pure-shrinkage solution is to place exchangeable double-exponential priors on the regression coefficients, a tactic which often goes by the name of “the Bayesian LASSO” (Park and Casella, 2008).

Additionally, we also review some recent developments about shrinkage and variable selection that are not explicitly computational in nature. We include these results in an attempt to give a current picture of the “state of the art” for Bayesian linear modeling.

## 2 BAYESIAN LINEAR MODELS

### 2.1 Notation

Given a vector  $Y$  of  $n$  responses and an  $n \times p$  design matrix  $X$ , suppose we wish to select a subset of  $k$  predictors, zeroing out the remaining  $p - k$  coefficients. This yields a sparse linear model of the form

$$Y_i = \alpha + X_{ij_1}\beta_{j_1} + \dots + X_{ij_k}\beta_{j_k} + \epsilon_i, \tag{1}$$

for some  $\{j_1, \dots, j_k\} \subset \{1, \dots, p\}$ , where  $\epsilon_i \stackrel{iid}{\sim} N(0, \phi^{-1})$ .

We follow the convention of treating the intercept  $\alpha$  differently, since all models will include this term. Let  $H_0$  denote the null model with only an intercept, and let  $H_F$  denote the full model with all covariates included. The full model thus has parameter vector  $\boldsymbol{\theta}' = (\alpha, \boldsymbol{\beta}')$ ,  $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_p)'$ .

Each model  $H_\gamma$  is indexed by a binary vector  $\gamma$  of length  $p$  indicating a set of  $k_\gamma \leq p$  nonzero regression coefficients  $\beta_\gamma$ :

$$\gamma_i = \begin{cases} 0 & \text{if } \beta_i = 0 \\ 1 & \text{if } \beta_i \neq 0. \end{cases}$$

In Bayesian model selection,  $\gamma$  itself is a random variable that takes values in the discrete space  $\{0, 1\}^p$ , which has  $2^p$  members. Inference relies upon the prior probability of each model,  $p(H_\gamma)$ , along with the marginal likelihood of the data under each model:

$$f(\mathbf{Y} | H_\gamma) = \int f(\mathbf{Y} | \boldsymbol{\theta}_\gamma, \phi) \pi(\boldsymbol{\theta}_\gamma, \phi) d\boldsymbol{\theta}_\gamma d\phi, \quad (2)$$

where  $\pi(\boldsymbol{\theta}_\gamma, \phi)$  is the prior for model-specific parameters. These together define, up to a constant, the posterior probability of a model:

$$p(H_\gamma | \mathbf{Y}) \propto p(H_\gamma) f(\mathbf{Y} | H_\gamma). \quad (3)$$

Let  $\mathbf{X}_\gamma$  denote the columns of the full design matrix  $\mathbf{X}$  given by the nonzero elements of  $\gamma$ , and let  $\mathbf{X}_\gamma^*$  denote the concatenation  $(\mathbf{1} \ \mathbf{X}_\gamma)$ , where  $\mathbf{1}$  is a column of ones corresponding to the intercept  $\alpha$ . For simplicity, assume that all covariates have been centered so that  $\mathbf{1}$  and  $\mathbf{X}_\gamma$  are orthogonal. Also assume that the common choice  $\pi(\alpha) = 1$  is made for the parameter  $\alpha$  in each model (see Berger et al., 1998, for a justification of this choice of prior).

Often all models will have small posterior probability, in which case more useful summaries of the posterior distribution are quantities such as the posterior inclusion probabilities of the individual variables:

$$w_i = \Pr(\gamma_i \neq 0 | \mathbf{Y}) = \sum_{\gamma} 1_{\gamma_i=1} \cdot p(H_\gamma | \mathbf{Y}). \quad (4)$$

These quantities also define the median-probability model, which is the model that includes those covariates having posterior inclusion probability of at least 1/2 (Barbieri and Berger, 2004).

## 2.2 Choosing Priors for Variable Selection

An extensive body of literature confronts the difficulties of Bayesian model choice in the face of weak prior information. These difficulties arise due to the obvious dependence of the marginal likelihoods in (2) upon the choice of priors for model-specific parameters. In general one cannot use improper priors on these parameters, since this leaves the resulting Bayes factors defined only up to an arbitrary multiplicative constant.

One class of methods for dealing with this issue involves training a noninformative prior using some function of the data, and then using the remaining data to compute Bayes factors under the induced family of prior distributions. This class includes the fractional Bayes factors of O’Hagan (1995) and the intrinsic Bayes factors of Berger and Pericchi (1996). An extensive discussion can be found in Berger and Pericchi (2001). Another promising recent method due to Ray et al. (2007) known as PBIC offers a default specification in terms of the principal components of the observed information matrix, and seems to offer an interesting alternative to the well known Bayesian information criterion (Schwarz, 1978), which can also be used to compute a set of pseudo-marginal likelihoods.

Other authors have sidestepped this problem by defining default proper priors that are appropriate for model selection and that explicitly aim to minimize the effect of the prior. One such example is the  $g$ -prior and its robust variants, where

$$(\beta_\gamma \mid g, \phi) \sim N \left\{ 0, \frac{g}{\phi} (\mathbf{X}_\gamma^t \mathbf{X}_\gamma)^{-1} \right\},$$

and where  $g$  is either chosen outright, given a prior, or estimated by marginal maximum likelihood.

The existence of simple expressions for marginal likelihoods has made the use of  $g$ -priors very popular. They can also be defended on foundational Bayesian grounds, since they automatically adjust the predictive distribution of a model to account for observed co-linearities in the variables (precisely the kind of behavior one would expect from a carefully done subjective elicitation).

Additionally, some recent authors have overcome one of the major problems of  $g$ -priors—namely, a type of unsettling behavior known as the “information paradox.” It turns out that robustifying the  $g$ -prior by giving it heavier-than-normal tails seems to solve this problem. Moreover, it does so in a way that does not make marginal likelihoods all that much more difficult to compute. Key references here are Zellner and Siow (1980), Zellner (1986), George and Foster (2000), and Liang et al. (2008). Another overview of  $g$ -type priors can be found in the appendix of Scott and Berger (2008).

What of the prior probabilities for models themselves? One might reasonably consider a set of subjective prior model probabilities in smaller problems. But the complexity of such an elicitation means that default methods must be developed as a practical matter for high-dimensional problems, or when the appearance of objectivity is important. In such cases, there seems to be wide agreement surrounding the use of so-called “variable selection priors,” where the  $p$ -dimensional vector  $\gamma$  is assumed to arise as a sequence of exchangeable Bernoulli trials with common success probability  $w$ .

In such cases, we find it natural to think of specifying prior model probabilities as an opportunity to apportion mass across model space in a way that solves the implicit problem of multiple hypothesis testing posed by variable selection. The

key intuition when using these priors is to let the data estimate  $w$ . This yields an automatic penalty for multiple testing, in that the introduction of spurious covariates will cause the posterior mass of  $w$  to concentrate near 0, making it harder for all variables to overcome the increasingly strong prior belief in their irrelevance (Scott and Berger, 2006). George and Foster (2000) propose estimating  $w$  by empirical-Bayes methods, but they note that this can prove computationally overwhelming in large problems with nonorthogonal design. Cui and George (2008) consider the fully Bayesian specification, whereby  $w$  is marginalized away before computing the posterior probability of a model. Finally, Scott and Berger (2008) offer some theoretical and numerical comparisons of the empirical-Bayes and fully Bayes approaches. They show that the fully Bayes solution offers an automatic improvement over empirical Bayes, in that it can avoid a particular form of degeneracy that arises when the empirical-Bayes solution collapses to the boundary of the parameter space.

### 3 ALGORITHMS FOR VARIABLE SELECTION AND SHRINKAGE

#### 3.1 Traditional MCMC

Computational algorithms for variable selection took flight beginning with the seminal work of George and McCulloch (1993) and followed by, among others, Geweke (1996), Clyde et al. (1996), and George and McCulloch (1997). These algorithms construct a Markov chain to simulate a sequence  $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \dots, \boldsymbol{\gamma}^{(T)}$  such that

$$\boldsymbol{\gamma}^{(t)} \xrightarrow{\mathcal{D}} p(H_{\boldsymbol{\gamma}} | \mathbf{Y})$$

as  $t \rightarrow \infty$ . The majority of these algorithms assume conjugate prior distributions to implement a Gibbs sampler (Gelfand and Smith, 1990) over the model space, since these allow marginal likelihoods to be computed in closed form. Several algorithms, however, allow non-conjugate priors to be used by employing Metropolis proposals (see, e.g., Madigan and York, 1995). In all cases, the inclusion probabilities (4) are estimated using the simulated  $\boldsymbol{\gamma}$  sequence, with,

$$\hat{w}_i = \frac{1}{T} \sum_{t=1}^T \mathbf{1}_{\gamma_i^{(t)}=1}. \quad (5)$$

An eloquent overview of these methods is given in Clyde and George (2004), which also includes a much more comprehensive list of references.

Due to their intuitive construction and ease of implementation, MCMC techniques for variable selection surged in popularity during the 1990's and at the turn of the century. In recent years, however, variable-selection techniques based upon traditional Markov-chain algorithms have come under scrutiny for a few reasons. First, for large  $p$ , the posterior distribution  $p(H_{\boldsymbol{\gamma}} | \mathbf{Y})$  is highly multi-

modal, and there are no trustworthy diagnostics that can effectively recognize a lack of Markov-chain convergence in such complex situations. The usual “rules of thumb” for MCMC, as we shall see on the examples later in the paper, can lead one badly astray when assessing convergence.

Second, many years of testing and implementation of such algorithms have shown that for a finite (and computationally practical) run time  $T$ , the chain often completely misses large modes in the model space. This potentially renders (5) a poor estimator of  $w_i$ . For many researchers that have studied this issue, it is very difficult to understand how a procedure can correctly estimate marginal distributions when it misses significant modes of the joint distribution. Even now, despite years of research on computational approaches for variable selection, it is simply not known whether the estimated inclusion probabilities that arise from MCMC on large problems are even approximately correct.

Third, it is very unlikely that the Markov chain will visit any model frequently enough to allow model probabilities to be estimated by frequency of occurrence in the Monte Carlo sample. In fact, in large problems, it will almost always be the case that all models (even the best one) will have posterior probabilities significantly smaller than  $1/T$ , which is the smallest nonzero model probability that can arise from an MCMC of length  $T$ .

### 3.2 Stochastic Search Algorithms for Variable Selection

For these and other reasons, some researchers have become skeptical of “vanilla” MCMC, and the popularity of these techniques as an active research area has dwindled. These older techniques have, however, paved the way for the emergence of newer stochastic-search (SS) algorithms, which focus on rapidly discovering models with high posterior probability. These algorithms use the information from previously visited models, such as estimated inclusion probabilities, to guide the search over the model space.

MCMC can, of course, be viewed as a form of stochastic search. But the SS algorithms discussed here pay little, if any, attention to the goal of converging to the posterior distribution  $p(H_\gamma | \mathbf{Y})$ . Rather, SS algorithms focus on finding the models with the highest posterior probability. Their output is simply a list of models visited, together with a score for each one—typically an unnormalized posterior probability. There is no sense which the estimated inclusion probabilities “converge” to the true ones, unless all of the models are eventually enumerated.

A simple SS algorithm proposed by Berger and Molina (2005), for example, uses online estimates of posterior model and inclusion probabilities to orient the search. Let  $p^{(t)}(H_\gamma | \mathbf{Y})$  and  $w_i^{(t)}$  be the estimates of  $p(H_\gamma | \mathbf{Y})$  and  $w_i$  at the  $t^{\text{th}}$  iteration of the SS algorithm, respectively. At iteration  $t$ , the algorithm proceeds by:

1. Resampling one of the  $t - 1$  previously sampled models in proportion to their estimated probabilities  $p^{(t)}(H_\gamma | \mathbf{Y})$ , and setting this to be the current model.
2. Flipping a coin to decide whether to add or delete a variable to the current model.
3. Adding [removing] variable  $i$  with probability  $(w_i^{(t)} + \epsilon)/(1 - w_i^{(t)} + \epsilon)$  [or  $(1 - w_i^{(t)} + \epsilon)/(w_i^{(t)} + \epsilon)$  in the case of a deletion], where  $\epsilon > 0$  is small and bounds  $w_i^{(t)}$  away from 0 or 1.

Berger and Molina (2005) suggest updating  $p^{(t)}(H_\gamma | \mathbf{Y})$  via a path-based estimate of the Bayes factors between models. Certainly the algorithm cannot explore all  $2^p$  models, but the hope is that a majority of visited models will have high posterior probability.

Many stochastic-search algorithms have this general flavor. The key ingredient to visiting good models seems to be to use the inclusion probabilities to guide the search—an approach that also works in far more general classes of models and features. For example, Scott and Carvalho (2008) propose a SS algorithm called FINCS (feature-inclusion stochastic search). This algorithm, which builds upon the insight of Berger and Molina (2005) regarding the importance of the inclusion probabilities, interweaves local moves (adding or deleting a variable from  $\gamma^{(t)}$ ), resampling moves (selection from among one of  $\gamma^{(1)}, \dots, \gamma^{(t-1)}$ ), and global moves that attempt to avoid getting stuck in local modes in model space. Their application of the algorithm is to Gaussian graphical models, but the approach is in principle quite straightforward to use in linear models, as well.

A SS algorithm of a slightly different nature was proposed by Hans et al. (2007), and is known as shotgun stochastic search (SSS). Hans et al. (2007) consider constructing a neighborhood of models around  $\gamma^{(t)}$  denoted by  $\partial\gamma^{(t)} = \{\gamma_+^{(t)}, \gamma_0^{(t)}, \gamma_-^{(t)}\}$  where  $\gamma_+^{(t)}, \gamma_0^{(t)}, \gamma_-^{(t)}$  is the set of all models which add, replace, or remove one element from  $\gamma^{(t)}$ , respectively. Each model in  $\partial\gamma^{(t)}$  is given a “score” (e.g. AIC, BIC, or a posterior probability), and a set  $\mathcal{S}^{(t)}$  is adapted to contain the  $B$  highest scoring models of  $\{\partial\gamma^{(t)}, \mathcal{S}^{(t-1)}\}$  such that  $\mathcal{S}^{(T)}$  contains the  $B$  best models after  $T$  iterations. To iterate the algorithm,  $\gamma^{(t+1)}$  is sampled from  $\partial\gamma^{(t)}$  proportional to the assigned scores. (Obviously, the SSS algorithm is computationally demanding and works best in a parallel computing environment, which it was designed to exploit.)

Clyde et al. (2009) astutely observe that for models with tractable marginal likelihoods, resampling a model provides no additional information for estimating posterior model probabilities. They go on to develop a Bayesian adaptive sampling (BAS) algorithm which samples without replacement from the  $2^p$  models. This too is accomplished by sampling models one variable at a time in a manner

that is guided by the estimated inclusion probabilities. If  $\mathcal{S}^{(t)}$  is the set of sampled models, then the estimated inclusion probability for the  $i^{\text{th}}$  variable is given by

$$\hat{w}_i^{(t)} = \frac{\sum_{\gamma \in \mathcal{S}^{(t)}} p(\mathbf{Y} | H_\gamma) \gamma_i}{\sum_{\gamma \in \mathcal{S}^{(t)}} p(\mathbf{Y} | H_\gamma)}, \quad (6)$$

and  $\hat{w}_i^{(t)} \rightarrow w_i$  as  $t \rightarrow 2^p$  because  $\mathcal{S}^{(t)}$  becomes the set of all  $2^p$  models. Sampling without replacement is ensured by subtracting the mass of model  $\gamma^{(t)}$  from the total mass of  $\pi(\gamma | \mathbf{Y})$ .

### 3.3 Adaptive MCMC Algorithms for Variable Selection

Similar to SS algorithms, the key idea of adaptive MCMC (AMCMC) is to inform and adapt the proposal distribution of a Metropolis-Hastings algorithm using past draws. Specifically, if  $X^{(1:t)} = \{X^{(i)} : i = 1 \dots, t\}$  is the set of realizations of the Markov chain  $X^{(t)}$  up to time up time  $t$ , then AMCMC would adapt the proposal distribution  $q(X^{(t)}, \cdot; \psi^{(t)})$  iteratively by adapting the parameter vector  $\psi^{(t)} = f(X^{(1:t)})$  for some function  $f$ .

As a simple example, suppose that the proposal density is  $q(X^{(t)}, \cdot; \psi^{(t)}) = N(\cdot; X^{(t)}, \sigma^{(t)})$ . Then an AMCMC algorithm could adapt  $\psi^{(t)} = \sigma^{(t)}$  iteratively via the update equation  $\psi^{(t)} = \sqrt{\text{Var}(X^{(1:t)})}$ . While this is a simple example, it illustrates the appeal of AMCMC in that the tuning of the proposal distribution is done automatically.

Because AMCMC algorithms use all past states  $X^{(1:t)}$  to construct the proposal distribution (i.e. estimate  $\psi^{(t)}$ ), the resulting algorithms no longer satisfy the Markov property: the past and future are no longer conditionally independent, given the present. Nevertheless, due to the recent theoretical work of, for example, Haario et al. (2001), Atchade and Rosenthal (2005), Andrieu and Moulines (2006), Andrieu and Atchadé (2007), Roberts and Rosenthal (2007), and Atchadé et al. (2009), simple and intuitive conditions have been established which, if met, guarantee that an AMCMC algorithm will converge to the desired posterior distribution. Using these conditions, practically useful algorithms have emerged for a variety of models and situations. These include Haario et al. (2001), Haario et al. (2005), Roberts and Rosenthal (2009), Pasarica and Gelman (2009), and Craiu et al. (2009).

Recently, some AMCMC methods for variable selection have begun to emerge. One of the first was proposed by Nott and Kohn (2005), who made clever use of the fact that  $Pr(\gamma_i = 1 | \gamma_{i^c}, \mathbf{Y}) = \mathbb{E}(\gamma_i | \gamma_{i^c}, \mathbf{Y})$ , where  $\gamma_{i^c} = \{\gamma_j \in \gamma : j \neq i\}$ . Specifically, Nott and Kohn (2005) adaptively estimate  $\bar{\gamma}^{(t)} = t^{-1} \sum_i \gamma^{(i)}$  and  $\Gamma^{(t)} = \text{Cov}(\gamma | \mathbf{Y})$  at each step of the AMCMC algorithm. They do so using the

best linear unbiased estimator of  $\mathbb{E}(\gamma_i \mid \boldsymbol{\gamma}_{i^c}, \mathbf{Y})$ ,

$$\widehat{\mathbb{E}}(\gamma_i \mid \boldsymbol{\gamma}_{i^c}, \mathbf{Y}) = \bar{\gamma}_i + \boldsymbol{\Gamma}_{i,i^c}^{(t)} \left[ \boldsymbol{\Gamma}_{i^c,i^c}^{(t)} \right]^{-1} (\boldsymbol{\gamma}_{i^c} - \bar{\boldsymbol{\gamma}}_{i^c}), \quad (7)$$

as a proposal distribution in the MCMC algorithm, where  $\boldsymbol{\Gamma}_{i,i^c}^{(t)}$  is the  $i^{\text{th}}$  row of  $\boldsymbol{\Gamma}^{(t)}$  with the  $i^{\text{th}}$  column removed, and  $\boldsymbol{\Gamma}_{i^c,i^c}^{(t)}$  is  $\boldsymbol{\Gamma}^{(t)}$  with the  $i^{\text{th}}$  row and column removed. Using (7) as a proposal distribution, variables with a high *conditional* inclusion probabilities are frequently added to the model.

The algorithm of Nott and Kohn (2005) uses conjugate prior distributions such that the coefficients and error precision can be integrated out, allowing a closed-form expression for  $f(\mathbf{Y} \mid H_\gamma)$ . An alternative AMCMC algorithm proposed by Ji and Schmidler (2009) use a point-mass mixture prior for the coefficients, e.g.

$$p(\beta_i) = (1 - w)\delta_0(\beta_i) + wN(\beta_i; 0, s_i^2), \quad (8)$$

where  $\delta_0(\cdot)$  is the Dirac measure at 0 and  $s_i^2$  is a known prior variance. Ji and Schmidler (2009) then consider adapting proposal distributions of the form,

$$q(\beta_i^{(t)}, \cdot; \boldsymbol{\psi}^{(t)}) = \lambda q_0(\cdot; \tilde{\boldsymbol{\psi}}) + (1 - \lambda) \left[ (1 - \omega^{(t)})\delta_0(\cdot) + \omega^{(t)}N(\cdot; \widehat{\boldsymbol{\beta}}_i^{(t)}, \widehat{\boldsymbol{\Sigma}}_i^{(t)}) \right], \quad (9)$$

where  $0 < \lambda < 1$  is fixed and known, and  $q_0(\cdot; \tilde{\boldsymbol{\psi}})$  is fixed (non-adaptive) to ensure that the bounded convergence condition in Theorem 13 of Roberts and Rosenthal (2007) is satisfied. Ji and Schmidler (2009) then use the stochastic approximation algorithm of Robbins and Monro (1951) to develop an adaptive scheme for  $\omega^{(t)}$ ,  $\widehat{\boldsymbol{\beta}}_j^{(t)}$ , and  $\widehat{\boldsymbol{\Sigma}}^{(t)}$  which minimizes the Kullback-Leibler (KL)-divergence between the target distribution  $\pi(\boldsymbol{\beta}_j \mid \mathbf{Y})$  and the proposal distribution (9).

One potential limitation of this algorithm is that it depends upon being able to write an exchangeable joint prior for the regression coefficients in a given model, where  $\beta_i \sim w \cdot p(\beta_i) + (1-w) \cdot \delta_0$  as in (8). This restriction excludes the possibility of using *g*-like priors, since these cannot be expressed using an exchangeable model for each coefficient. Since there are strong (non-computational) reasons to prefer *g*-like priors for variable selection in situations with non-orthogonal designs, this limitation may be a significant one.

### 3.4 Shrinkage-based alternatives

All of the models discussed so far place nonzero probability mass upon the hypothesis that each coefficient  $\beta_i$  is zero. As we have seen, this results in a combinatorial explosion in the number of discrete models that must be considered, leading to a very difficult problem in stochastic computation.

Recently, many researchers have become interested in an alternative approach

based upon “pure shrinkage” priors, which do not place positive probability at zero. There is, of course, an established tradition of evaluating such priors on foundational Bayesian grounds (see, e.g., Pericchi and Smith, 1992). Yet much of the more recent activity has arisen as a Bayesian rejoinder to the neo-classical literature on penalized least squares, which offers a very different perspective on variable selection. Indeed, many well-studied priors are enjoying newfound prosperity in their second careers as “penalty functions,” which yield solutions that can be interpreted as posterior modes.

These priors are often in the family of multivariate scale mixtures of normals, which is very general and has many nice analytical properties:

$$\begin{aligned} (Y \mid \beta, \sigma^2) &\sim N(X\beta, \sigma^2 I) \\ (\beta_j \mid \lambda_j, \tau, \sigma^2) &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\ \lambda_j &\sim g(\lambda_j) \\ \tau &\sim h(\tau). \end{aligned}$$

The  $\lambda_j$ 's are known as the “local” shrinkage parameters, while  $\tau$  is known as the “global” shrinkage parameter.

The following list is by no means comprehensive, but gives a sense of the strong level of activity in this area:

1. The horseshoe prior of Carvalho et al. (2008) assumes a half-Cauchy prior on the local scales,  $\lambda_j \sim C^+(0, 1)$ , which is equivalent to an  $F(1, 1)$  prior on the local variances  $\lambda_j^2$ . Polson and Scott (2009) generalize this prior to a wider class of hypergeometric–beta mixtures, while Scott (2009) proposes two methods for fitting models in this family: one based on importance sampling, and an alternative MCMC algorithm that involves a slice-sampling step for the local shrinkage parameters.
2. The Student- $t$  prior is defined by an inverse-gamma mixing density,  $\lambda_j^2 \sim \text{IG}(\xi/2, \xi\tau^2/2)$ . Tipping (2001) uses this model for sparsity by finding posterior modes under the assumption that  $\xi \rightarrow 0$ .
3. The double-exponential prior uses an exponential mixing density:  $p(\lambda_j^2 \mid \tau^2) \propto \exp\{\lambda_j^2/2\tau^2\}$ . The standard Markov-chain Monte Carlo algorithm for working with this model is from Carlin and Polson (1991), and uses the fact that the local variance parameters are conditionally inverse-Gaussian, given the data and other parameters. More recently, Park and Casella (2008), along with many others such as Hans (2009) and Gramacy and Pantaleo (2009), have revitalized interest in this prior as a Bayesian alternative to the LASSO (Tibshirani, 1996).
4. The normal–Jeffreys prior has been studied by Figueiredo (2003) and Bae and Mallick (2004). This improper prior is induced by placing Jeffreys’

prior upon each variance term,  $p(\lambda_j^2) \propto 1/\lambda_j^2$ , leading to  $p(\beta_j) \propto |\beta_j|^{-1}$  independently.

5. The normal–exponential–gamma family of priors proposed by Griffin and Brown (2005) is also based upon the exponential mixing density, but uses a  $\text{Ga}(c, d^2)$  density rather than an inverse-gamma for the global scale term  $\tau$ . The two hyperparameters allow control over tail weight ( $c$ ) and scale ( $d$ ). This leads to

$$p(\lambda_j^2) = \frac{c}{d^2} \left(1 + \frac{\lambda_j^2}{d^2}\right)^{-(c-1)}.$$

Many of these priors are implemented in the R package `monomvn` (Gramacy, 2009), which we use in later sections to fit the shrinkage-based methods. A discussion of some general principles to help guide the choice of prior can be found in Carvalho et al. (2008), who compare many of the above possibilities at great length. Their conclusion is that, in order to be appropriate for sparse problems, the prior for  $\lambda_j$  should have positive density at zero, and should decay no faster than  $\lambda_j^{-2}$ . (These same guidelines also apply to the prior on  $\tau$ .)

To be sure, pure-shrinkage solutions can never provide a truly sparse solution, in the sense that they will never allocate positive posterior probability at zero. Nonetheless, there is a growing body of empirical evidence to suggest that it *is* possible to use pure-shrinkage priors to get estimates and predictions very close to those that arise under Bayesian model averaging. This is an active and fast-moving area of research, and the exciting possibility of “BMA mimicry” using shrinkage priors is just one of many open problems here.

## 4 EXAMPLES

In this section, the approaches described above are evaluated on three examples: a very simple, simulated orthogonal problem; a data set on the long-term economic growth rates of 88 countries; and a data set of daily maximum ozone measurements near Los Angeles. We address four questions that are relevant to comparing MCMC and stochastic search/adaptive sampling, the two general classes of variable-selection algorithms that have been considered here:

1. Does either class of methods systematically find better models?
2. Do the classes systematically differ in their estimates of inclusion probabilities?
3. Does either class yield better out-of-sample performance?
4. How do pure-shrinkage solutions compare to full-blown model averaging?

## 4.1 Orthogonal Simulation Study

Our first experiment is designed to test the algorithms in a situation where the model space is too large to enumerate, but where everything else remains as simple as possible. Hence we construct an orthogonal problem with no unknown hyper-parameters, where all true inclusion probabilities are known exactly, and where the identity of the top model is known.

Specifically, we let

$$Y_i \stackrel{iid}{\sim} N(\mu_i \gamma_i, \sigma^2) \quad (10)$$

for  $i = 1, \dots, n$ . Here  $\gamma_i$  is either 1 or 0, designating signal or noise. We chose  $n = 50$  and  $\sigma^2 = 1$ , and we assume that the nonzero means follow  $\mu_i \sim N(0, 1)$ , and that  $Pr(\gamma_i = 1) = 0.5$  independently for all  $i$ . Even though the full model space has  $2^{50}$  members and is too big to enumerate, the structure of the problem allows inclusion probabilities to be computed exactly. The marginal likelihood of the data under a model configuration  $\gamma$  is

$$f(\mathbf{Y} | H_\gamma) = \prod_i N(Y_i | 0, 1 + \gamma_i),$$

where  $N(x | m, v)$  is the normal p.d.f. with mean  $m$  and variance  $v$  evaluated at  $x$ . Meanwhile, under this simple design, the true inclusion probabilities are

$$w_i = \frac{N(Y_i | 0, 2)}{N(Y_i | 0, 1) + N(Y_i | 0, 2)}, \quad (11)$$

and the highest posterior probability model is the median probability model.

We actually simulated three data sets under (10) with low, medium, and high signal-to-noise (STN) ratios. The low-STN data set takes  $\mu_i = i$  for  $i = 1, \dots, 5$ ; the medium STN data set takes  $\mu_i = i/2$  for  $i = 1, \dots, 10$ ; and the high STN data set takes  $\mu_i = i/5$  for  $i = 1, \dots, 25$ . (All other means are set to zero.)

For each simulated data set, we attempted to reconstruct the posterior distribution for  $\gamma$  using the stochastic-search algorithm of Berger and Molina (2005) (FINCS), the AMCMC algorithm of Nott and Kohn (2005), and the SSVS algorithm of George and McCulloch (1993). Each MCMC was run for  $T = 5000$  iterations after discarding an initial 500 iterations for burn-in, while FINCS was run for 250,000 iterations. (These numbers mean that each algorithm evaluated the same number of marginal likelihoods, making the comparison a fair one.) Table 1 displays the sum of absolute errors for inclusion probabilities  $SAE_w = \sum_i |w_i - \hat{w}_i|$  for the three data sets, and Figure 1 displays corresponding boxplots of  $\log f(\mathbf{Y} | \gamma)$  of the top visited models.

Additionally, the full set of inclusion probabilities for the medium STN experiment can be found in Table 2. The table is a bit dense but repays close inspection, since together with Figure 1 it tells a very interesting story. On the

Table 1: Sum of absolute error in inclusion probabilities for orthogonal simulation study.

Data	SSVS	FINCS	AMCMC
Low Density	0.284	12.213	0.384
Medium Density	0.221	10.135	0.394
High Density	0.208	8.391	0.372

one hand, the FINCS algorithm is quite poor at estimating inclusion probabilities compared to AMCMC or SSVS, even on this simple orthogonal problem. In particular, it seems to overestimate  $w_i$  for “good” variables, and to underestimate  $w_i$  for “bad” variables. (This was also true for the low- and high-STN data sets, though these tables are omitted.) This systematic bias is interesting but perhaps not too surprising: FINCS is not concerned with exploring all models, nor with re-constructing any marginal distributions.

Meanwhile, both SSVS and AMCMC get the inclusion probabilities essentially correct. Yet paradoxically, the explored models under FINCS have a higher marginal likelihood than those found under either AMCMC or FINCS. Indeed, FINCS finds dozens of models that are better than the single best one discovered by either SSVS or AMCMC. This fact is much harder to understand: how is it that, at least in this case, both MCMC methods are able to reconstruct the correct marginal distributions while missing large pockets of probability in the joint distribution from which all these marginals are derived?

#### 4.2 GDP Growth Data

We next ran a similar experiment on a real data set that was collected in an attempt to understand the determinants of long-term economic growth. Here  $Y$  is annualized GPD growth since 1960 for 88 countries, and  $X$  represents a battery of 67 possible socio-economic, political, and geographical predictors of growth. This data set has been previously analyzed by Fernandez et al. (2001), Sala-i Martin et al. (2004), and Ley and Steel (2007). We assume  $g$ -priors for the coefficients; unlike in the orthogonal problem, the true inclusion probabilities are unknown.

Surprisingly, a very different pattern emerged. Before, SSVS and AMCMC agreed (both with each other and with the truth), while FINCS disagreed despite visiting better models. On this problem, however, FINCS and AMCMC tend to agree with each other—though not perfectly—while SSVS disagrees with both of them. As Table 3 shows, this disagreement can be stark. For example, SSVS estimates the inclusion probability of the East Asian dummy variable to be 50%,

Table 2: True inclusion probabilities for the 50 simulated coefficients in the medium signal-to-noise-ratio configuration, along with estimates arising from three algorithms: stochastic search using inclusion probabilities (FINCS), Gibbs sampling over models (SSVS), and adaptive Markov-chain Monte Carlo (AMCMC). The results are rounded to two decimal places and are ordered by the absolute value of the observation  $Y_i$ .

Rank	Y	True $w_i$	FINCS	SSVS	AMCMC
1	5.98	1.00	1.00	1.00	1.00
2	4.94	1.00	1.00	1.00	1.00
3	4.30	1.00	1.00	1.00	1.00
4	3.80	0.99	1.00	0.99	0.99
5	3.38	0.97	1.00	0.97	0.97
6	3.10	0.95	1.00	0.94	0.95
7	2.75	0.90	0.99	0.90	0.90
8	-2.71	0.89	0.99	0.89	0.89
9	-2.14	0.74	0.95	0.75	0.74
10	-1.87	0.65	0.92	0.65	0.66
11	1.67	0.59	0.86	0.59	0.59
12	-1.63	0.58	0.75	0.57	0.58
13	-1.60	0.57	0.74	0.57	0.58
14	1.57	0.56	0.75	0.55	0.55
15	-1.55	0.55	0.69	0.55	0.55
16	1.44	0.52	0.55	0.52	0.53
17	1.42	0.52	0.69	0.51	0.52
18	-1.29	0.48	0.36	0.48	0.49
19	1.28	0.48	0.70	0.48	0.49
20	1.26	0.48	0.49	0.48	0.48
21	1.23	0.47	0.31	0.46	0.48
22	-1.23	0.47	0.43	0.46	0.48
23	1.13	0.45	0.22	0.44	0.44
24	1.03	0.43	0.14	0.42	0.42
25	-0.85	0.40	0.18	0.40	0.39
26	0.74	0.38	0.09	0.39	0.36
27	0.71	0.38	0.08	0.38	0.36
28	-0.69	0.37	0.08	0.38	0.37
29	-0.66	0.37	0.08	0.37	0.37
30	0.65	0.37	0.07	0.36	0.36
31	0.61	0.36	0.09	0.36	0.35
32	0.59	0.36	0.08	0.37	0.35
33	0.58	0.36	0.07	0.35	0.34
34	0.57	0.36	0.08	0.36	0.35
35	-0.55	0.36	0.16	0.35	0.35
36	0.50	0.36	0.06	0.36	0.35
37	-0.50	0.36	0.08	0.36	0.37
38	-0.46	0.35	0.07	0.34	0.35
39	0.42	0.35	0.10	0.36	0.34
40	-0.36	0.34	0.07	0.35	0.33
41	-0.33	0.34	0.06	0.34	0.34
42	0.26	0.34	0.06	0.33	0.32
43	0.22	0.34	0.07	0.34	0.33
44	-0.21	0.34	0.06	0.33	0.32
45	0.19	0.34	0.06	0.34	0.32
46	-0.18	0.34	0.07	0.33	0.32
47	-0.14	0.34	0.06	0.34	0.34
48	-0.07	0.33	0.06	0.34	0.31
49	0.02	0.33	0.05	0.33	0.33
50	-0.01	0.33	0.06	0.33	0.32

Table 3: Estimated inclusion probabilities for the top 50 (out of 67) variables in the GDP growth data set. Results are given for SSVS, FINCS, and AMCMC. AMCMC was replicated three times to ensure stability, with results displayed for all three runs.

Variable	SSVS	FINCS	AMCMC 1	AMCMC 2	AMCMC 3
Investment Price	0.98	1.00	1.00	1.00	1.00
GDP in 1960 (log)	0.97	1.00	1.00	1.00	1.00
Primary Schooling in 1960	0.94	0.99	1.00	1.00	1.00
Fraction Confucian	0.73	1.00	1.00	1.00	0.99
Fraction GDP in Mining	0.72	1.00	0.99	0.99	0.99
Public Investment Share	0.64	0.99	0.98	0.98	0.98
African Dummy	0.55	1.00	0.96	0.97	0.98
Fraction Buddhist	0.52	0.93	0.94	0.96	0.95
East Asian Dummy	0.50	0.02	0.04	0.03	0.03
Fraction Speaking Foreign Language	0.48	0.85	0.83	0.81	0.84
Life Expectancy in 1960	0.47	0.51	0.59	0.56	0.56
Fraction Muslim	0.43	0.12	0.16	0.14	0.13
Fraction of Tropical Area	0.41	0.13	0.13	0.16	0.12
Latin American Dummy	0.41	1.00	0.96	0.96	0.98
Population Density Coastal in 1960s	0.39	0.09	0.11	0.12	0.10
Population Density 1960	0.37	0.05	0.03	0.03	0.03
Real Exchange Rate Distortions	0.34	0.07	0.07	0.06	0.05
Nominal Gov. GDP Share 1960s	0.33	0.10	0.09	0.09	0.07
Gov. Consumption Share 1960s	0.31	0.30	0.24	0.35	0.29
Real Gov. GDP Share in 1960s	0.30	0.54	0.54	0.42	0.52
Revolutions and Coups	0.28	0.40	0.26	0.29	0.31
Fraction Catholic	0.27	0.07	0.04	0.05	0.04
Openess measure 1965-74	0.27	0.10	0.07	0.08	0.06
Fertility in 1960s	0.25	0.14	0.09	0.07	0.09
Hydrocarbon Deposits in 1993	0.24	0.04	0.02	0.03	0.02
Fraction Hindus	0.24	0.04	0.05	0.04	0.04
European Dummy	0.22	0.03	0.04	0.03	0.02
Ethnolinguistic Fractionalization	0.21	0.02	0.01	0.01	0.02
Outward Orientation	0.20	0.17	0.08	0.07	0.09
Fraction Protestants	0.20	0.02	0.01	0.01	0.01
Spanish Colony	0.02	0.03	0.03	0.03	0.02
Fraction Population In Tropics	0.20	0.06	0.05	0.04	0.04
Political Rights	0.20	0.02	0.01	0.01	0.01
Civil Liberties	0.20	0.04	0.02	0.02	0.02
Years Open 1950-94	0.20	0.02	0.01	0.01	0.01
Primary Exports 1970	0.20	0.05	0.03	0.03	0.03
Fraction Population Over 65	0.19	0.03	0.03	0.03	0.02
Colony Dummy	0.17	0.02	0.01	0.01	0.01
Air Distance to Big Cities	0.17	0.02	0.01	0.01	0.01
Higher Education 1960	0.17	0.02	0.01	0.01	0.01
Education Spending Share, 1960s	0.17	0.05	0.02	0.02	0.02
Socialist Dummy	0.17	0.06	0.02	0.03	0.03
Malaria Prevalence in 1960s	0.17	0.05	0.03	0.03	0.03
Capitalism	0.16	0.04	0.02	0.02	0.02
Population in 1960	0.16	0.03	0.01	0.01	0.01
Absolute Latitude	0.16	0.02	0.01	0.01	0.01
Fraction Land Near Navigable Water	0.16	0.03	0.01	0.02	0.01
Fraction Population Less than 15	0.16	0.03	0.01	0.01	0.01
British Colony Dummy	0.16	0.03	0.01	0.01	0.01
Landlocked Country Dummy	0.14	0.02	0.01	0.01	0.01

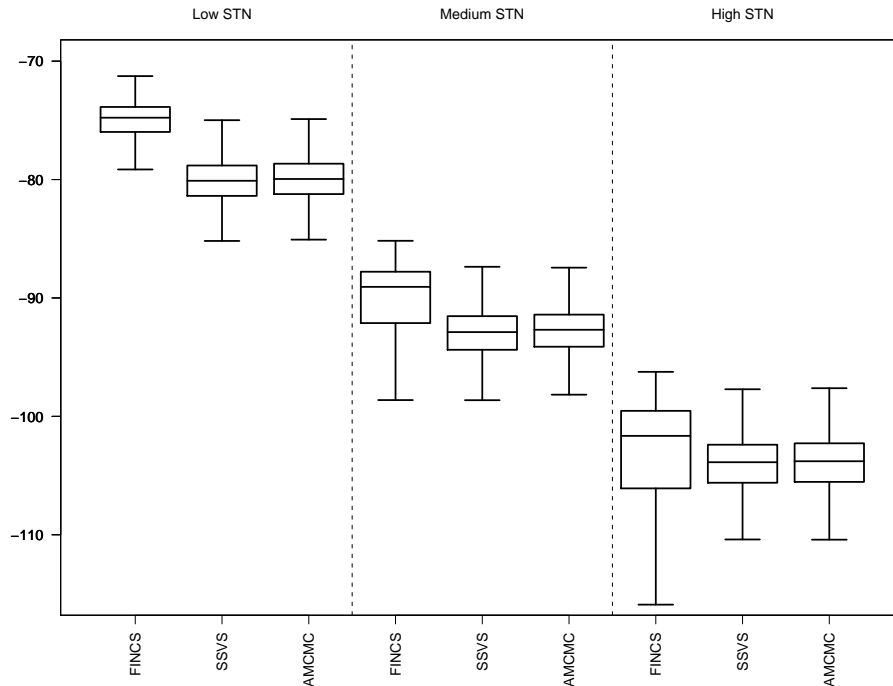


Figure 1:  $\log f(\mathbf{Y} | H_\gamma)$  of the explored models using FINCS, SSVS, and AMCMC for the low, medium, and high STN data sets.

while neither of the other methods estimate this probability to be larger than 4%.

Given a sufficient burn-in period, both SSVS and AMCMC are fairly stable from run to run. (The burn period tends to be quite long for AMCMC, but not untenably so.) This creates the illusion that each has independently converged to the posterior distribution. Yet at least one of them certainly has not, and it is impossible to know which one it is using existing tools.

A final fact worth noting is that, as before, SSVS fails to visit many high-probability models (Figure 2). Indeed, the cumulative posterior probability of all models discovered by SSVS is only 0.6% that of the top 10,000 models visited by FINCS.

### 4.3 Ozone Data and Out-of-Sample Performance

The ozone data set consists of  $n = 178$  daily measurements of the maximum ozone concentration near Los Angeles. This data set has become a standard benchmark in the regression literature, and has been recently analyzed by, among others, Casella and Moreno (2005), Berger and Molina (2005), and Liang et al.

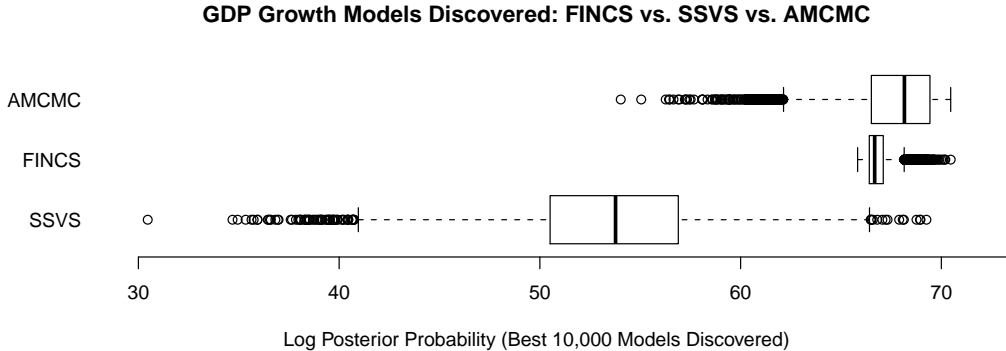


Figure 2: Log marginal likelihoods of models discovered by the three algorithms on the GDP growth example.

(2008). For this study, 10 atmospheric predictor variables are considered (see Casella and Moreno 2005 for a description), along with all squared terms and all 45 second-order interactions. This yields  $p = 65$  potential variables that could be included in the model. Enumerating all  $2^{65}$  models is impossible—to store all of the binary vectors on a computer would require 300 million terabytes of memory.

For this study, we performed 100 different “train–test” splits of the data set: a random sample of 134 data points were used to fit the model, with the remaining 44 used to compare out of sample predictive performance. The test subjects were: the AMCMC algorithm of Nott and Kohn (2005), the SSVS algorithm of George and McCulloch (1993), the BAS algorithm of Clyde et al. (2009), the horseshoe (HS) method of Carvalho et al. (2008), the Bayesian lasso, and the classical lasso. Zellner’s  $g$ -prior was used with  $g = n$  for the AMCMC, BAS, and SSVS algorithms.

Figure 3 displays box plots of the sum of predictive squared errors for the 100 repetitions,

$$SPSE = \sum_{i \in \mathcal{V}} (Y_i - \hat{Y}_i)^2,$$

where  $\mathcal{V}$  is set of indices of the 44 data points in the test data set, and  $\hat{Y}_i$  is one of either the model-averaged estimate of  $Y_i$  when using AMCMC, BAS, and SSVS; the posterior predictive mean when using the Bayesian lasso and the horseshoe; or the posterior predictive mode when using the lasso.

Each of the methods performed very similarly in terms of prediction, with median SPSE’s of 833.99 for AMCMC, 794.50 for SSVS, 847.27 for BAS, 837.51 for the Bayesian lasso, 821.56 for the HS, and 898.58 for the classical lasso. Clearly the between-sample variation is much larger than the between-method variation.

While the methods were similar at predicting, however, they differed greatly in

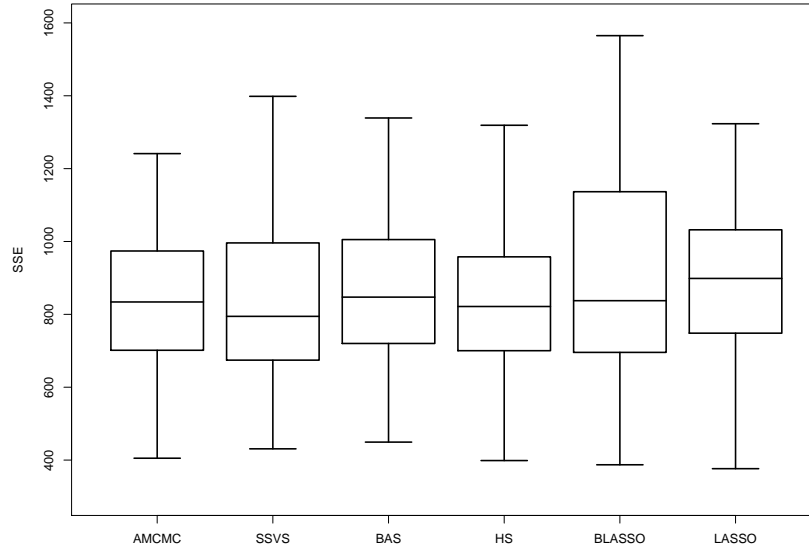


Figure 3: Box plots of sum of predictive squared errors for ozone cross-validation study.

average model size. Adaptive MCMC, with an average model size of 6.46, found the most parsimonious models; SSVS, with an average model size of 16.18, found the most complex models. Bayesian adaptive sampling and the classical lasso were intermediate, yielding average model sizes of 11.56 and 11.84, respectively. Thus, while each method is performing similarly in terms of SPSE, the models being chosen are quite different. It is particularly difficult to explain why this is the case for AMCMC and SSVS, since both algorithms are theoretically converging to the same posterior distribution.

Also noteworthy is that the pure-shrinkage solutions are competitive with Bayesian model-averaging in terms of out-of-sample predictions. It is not at all clear, however, how one would select a model or construct a measure of variable importance under pure-shrinkage priors. Sparse solutions based on the posterior mode are clearly dubious from a Bayesian point of view; except for the very rare case of a true “0–1” loss function, there is no deeper justification for choosing *any* point in  $\mathcal{R}^p$  with zero posterior probability, beyond a simple desire to induce sparsity.

## 5 FINAL REMARKS

Where do these results leave those interested in fitting a Bayesian linear model? We cannot, unfortunately, give an unqualified recommendation for any algorithm. We can only point to specific areas in which each one fails.

First, it is clear that if one's goal is to find models with high posterior probability, then stochastic search is preferred to either AMCMC or SSVS. This message emerges again and again from our simulation studies, and from those of other authors: existing MCMC methods in  $\gamma$  space simply miss too many good models.

Second, as a general matter, it remains unclear how one should compute posterior inclusion probabilities. Gibbs and AMCMC seem to reconstruct these probabilities in orthogonal settings, but no one knows whether this fidelity of reconstruction holds in high-dimensional, nonorthogonal problems. The experiment on the GDP-growth data suggests that it may not. Perhaps the best we can hope for at present is to estimate a set of *conditional* inclusion probabilities—conditioning on the set of models actually visited, and working hard to ensure that this set is a good one.

## REFERENCES

- Andrieu, C. and Atchadé, Y. F. (2007), “On the Efficiency of Adaptive MCMC Algorithms,” *Electronic Communications in Probability*, 12.
- Andrieu, C. and Moulines, E. (2006), “On the Ergodicity Properties of some Adaptive MCMC Algorithms,” *Annals of Applied Probability*, 16, 1462–1505.
- Atchadé, Y. F., Fort, G., Moulines, E., and Priouret, P. (2009), “Adaptive Markov chain Monte Carlo: Theory and Methods,” Tech. rep., University of Michigan.
- Atchade, Y. F. and Rosenthal, J. S. (2005), “On Adaptive Markov Chain Monte Carlo Algorithms,” *Bernoulli*, 11, 815–828.
- Bae, K. and Mallick, B. (2004), “Gene selection using a two-level hierarchical Bayesian model,” *Bioinformatics*, 20, 3423–30.
- Barbieri, M. and Berger, J. O. (2004), “Optimal predictive model selection,” *The Annals of Statistics*, 32, 870–897.
- Berger, J., Pericchi, L., and Varshavsky, J. (1998), “Bayes factors and marginal distributions in invariant situations,” *Sankhya, Ser. A*, 60, 307–321.
- Berger, J. O. and Molina, G. (2005), “Posterior model probabilities via path-based pairwise priors,” *Statistica Neerlandica*, 59, 3–15.
- Berger, J. O. and Pericchi, L. (1996), “The intrinsic Bayes factor for model selection and prediction,” *Journal of the American Statistical Association*, 91, 109–122.

- (2001), “Objective Bayesian methods for model selection: introduction and comparison,” in *Model Selection*, Beachwood, vol. 38 of *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, pp. 135–207.
- 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.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2008), “The horseshoe estimator for sparse signals,” Discussion Paper 2008-31, Duke University Department of Statistical Science.
- Casella, G. and Moreno, E. (2005), “Objective Bayesian Variable Selection,” Tech. rep., University of Florida.
- Clyde, M., Desimone, H., and Parmigiani, G. (1996), “Prediction via Orthogonalized Model Mixing,” *Journal of the American Statistical Association*, 91, 1197–208.
- Clyde, M. and George, E. I. (2004), “Model Uncertainty,” *Statistical Science*, 19, 81–94.
- Clyde, M., Ghosh, J., and Littman, M. (2009), “Bayesian Adaptive Sampling for Variable Selection,” Tech. rep., Duke University.
- Craiu, R. V., Rosenthal, J. S., and Yang, C. (2009), “Learn from Thy Neighbor: Parallel-Chain Adaptive MCMC,” *Journal of the American Statistical Association*, to appear.
- Cui, W. and George, E. I. (2008), “Empirical Bayes vs. fully Bayes variable selection,” *Journal of Statistical Planning and Inference*, 138, 888–900.
- Fernandez, C., Ley, E., and Steel, M. (2001), “Model Uncertainty in Cross-Country Growth Regressions,” *Journal of Applied Econometrics*, 16, 563–76.
- Figueiredo, M. (2003), “Adaptive sparseness for supervised learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–9.
- Gelfand, A. E. and Smith, A. F. M. (1990), “Sampling-Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, 85, 398–409.
- George, E. I. and Foster, D. P. (2000), “Calibration and empirical Bayes variable selection,” *Biometrika*, 87, 731–747.
- George, E. I. and McCulloch, R. (1993), “Variable selection via Gibbs sampling,” *Journal of the American Statistical Association*, 88, 881–889.
- George, E. I. and McCulloch, R. E. (1997), “Approaches for Bayesian Variable Selection,” *Statistica Sinica*, 7, 339–374.
- Geweke, J. (1996), “Variable selection and model comparison in regression.” in *Bayesian Statistics 5*, eds. Bernardo, J., Berger, J., Dawid, A. P., and Smith, A. F. M., Oxford University Press, pp. 609–620.
- Gramacy, R. and Pantaleo, E. (2009), “Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing,” Tech. rep., University of Cambridge, arXiv:0907.2135v1.

- Gramacy, R. B. (2009), *monomvn: Estimation for multivariate normal and Student-t data with monotone missingness*, R package version 1.7-3.
- Griffin, J. and Brown, P. (2005), “Alternative prior distributions for variable selection with very many more variables than observations,” Tech. rep., University of Warwick.
- Haario, H., Saksman, E., and Tamminen, J. (2001), “An Adaptive Metropolis Algorithm,” *Bernoulli*, 2, 223–242.
- (2005), “Componentwise Adaptation for High Dimensional MCMC,” *Computational Statistics*, 20, 265–273.
- Hans, C., Dobra, A., and West, M. (2007), “Shotgun stochastic search in regression with many predictors,” *Journal of the American Statistical Association*, 102, 507–516.
- Hans, C. M. (2009), “Bayesian Lasso Regression,” *Biometrika*, to appear.
- Ji, C. and Schmidler, S. C. (2009), “Adaptive Markov chain Monte Carlo for Bayesian Variable Selection,” Tech. rep., Duke University.
- Ley, E. and Steel, M. (2007), “Jointness in Bayesian variable selection with applications to growth regression,” *Journal of Macroeconomics*, 29, 476–493.
- Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008), “Mixtures of  $g$ -priors for Bayesian variable selection,” *Journal of the American Statistical Association*, 103, 410–23.
- Madigan, D. and York, J. (1995), “Bayesian graphical models for discrete data,” *International Statistical Review*, 63.
- Nott, D. and Kohn, R. (2005), “Adaptive Sampling for Bayesian Variable Selection,” *Biometrika*, 92.
- O’Hagan, A. (1995), “Fractional Bayes factors for model comparison,” *Journal of the Royal Statistical Society, Series B*, 57, 99–138.
- Park, T. and Casella, G. (2008), “The Bayesian Lasso,” *Journal of the American Statistical Association*, 103, 681–6.
- Pasarica, C. and Gelman, A. (2009), “Adaptively Scaling the Metropolis Algorithm using Expected Squared Jumped Distance,” *Statistica Sinica*, in press.
- Pericchi, L. R. and Smith, A. (1992), “Exact and Approximate Posterior Moments for a Normal Location Parameter,” *Journal of the Royal Statistical Society (Series B)*, 54, 793–804.
- Polson, N. G. and Scott, J. G. (2009), “Alternative global–local shrinkage rules using hypergeometric–beta mixtures,” Tech. Rep. 14, Duke University Department of Statistical Science.
- Ray, S., Berger, J. O., Bayarri, M., and Pericchi, L. R. (2007), “An Extended BIC for Model Selection,” Presentation at the Joint Statistical Meetings, Salt Lake City.
- Robbins, H. and Monro, S. (1951), “A Stochastic Approximation Method,” *Annals of Mathematical Statistics*, 22, 400–407.

- Roberts, G. O. and Rosenthal, J. S. (2007), “Coupling and Ergodicity of Adaptive MCMC,” *Journal of Applied Probability*, 44.
- (2009), “Examples of Adaptive MCMC,” *Journal of Computational and Graphical Statistics*, submitted.
- Sala-i Martin, X., Doppelhofer, G., and Miller, R. I. (2004), “Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach,” *American Economic Review*, 94, 813–835.
- Schwarz, G. (1978), “Estimating the dimension of a model,” *Annals of Statistics*, 6, 461–464.
- Scott, J. G. (2009), “Flexible Learning on the Sphere Via Adaptive Needlet Shrinkage and Selection,” Tech. Rep. 09, Duke University Department of Statistical Science.
- Scott, J. G. and Berger, J. O. (2006), “An exploration of aspects of Bayesian multiple testing,” *Journal of Statistical Planning and Inference*, 136, 2144–2162.
- (2008), “Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem,” Discussion Paper 2008-10, Duke University Department of Statistical Science.
- Scott, J. G. and Carvalho, C. M. (2008), “Feature-inclusion stochastic search for Gaussian graphical models,” *Journal of Computational and Graphical Statistics*, 17.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *J. Royal. Statist. Soc B.*, 58, 267–88.
- Tipping, M. (2001), “Sparse Bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, 1, 211–44.
- Zellner, A. (1986), “On assessing prior distributions and Bayesian regression analysis with  $g$ -prior distributions,” in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Elsevier, pp. 233–243.
- Zellner, A. and Siow, A. (1980), “Posterior odds ratios for selected regression hypotheses,” in *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia*, pp. 585–603.